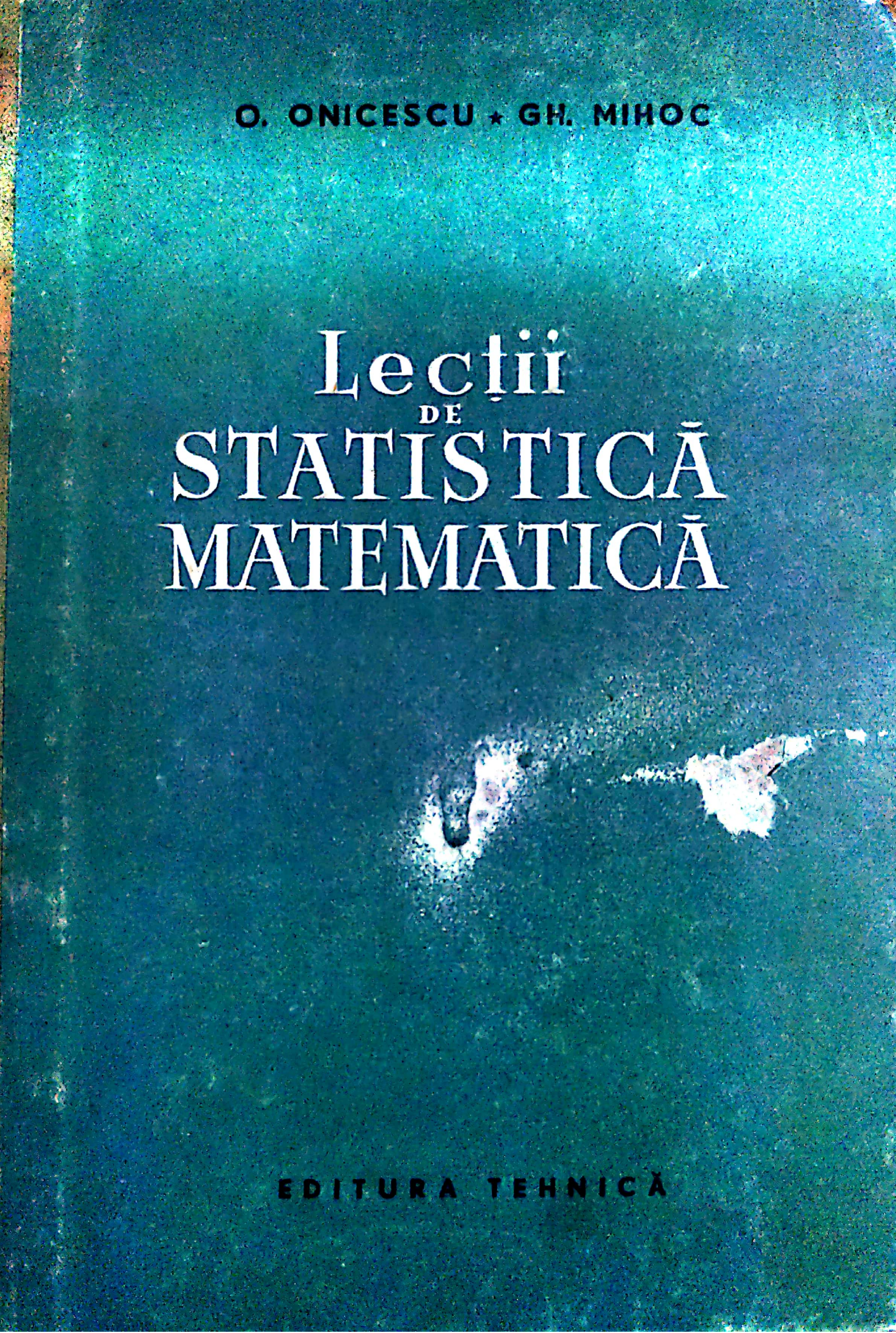
O. ONICESCU ★ GH. MIHOC

Lecții DE STATISTICĂ MATEMATICĂ



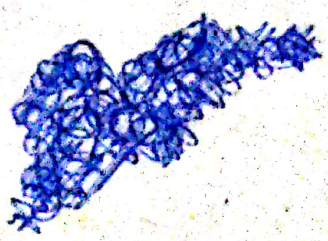
EDITURA TEHNICĂ

O. ONICESCU ★ GH. MIHOC

Lecții DE STATISTICĂ MATEMATICĂ

Redactate de: V. NEICOV și D. FIRESCU

EDITURA TEHNICĂ
BUCUREȘTI 1958



Scanned with OKEN Scanner

PREFAȚĂ

Cartea pe care o prezentăm redă conținutul unor lecții ținute la Facultatea de Matematică și Fizică din București în ultimii ani și reprezintă o introducere în domeniul foarte vast al Statisticii matematice. Această ramură a științelor matematice s-a dezvoltat pe fondul teoriilor statistice embrionare ale diferitelor științe ale naturii și ale tehnicii. Insă numai în ultimii 30 de ani, după ce teoria probabilităților constituită ea însăși ca știință matematică autonomă a clădit un fundament solid conceptelor și metodelor proprii ale statisticei, aceasta din urmă a luat un avînt, necontenit îmboldit de cerințele celorlalte științe.

O prezentare completă, sistematică într-un singur volum a principalelor teorii statistice pare astăzi prematură. Fiecare încercare de a da o astfel de prezentare este repede depășită chiar pentru domenii esențiale ale acestei științe — așa cum se vede din exemplul celor cîteva foarte bune cărți apărute în literatura țărilor în care statistica matematică este în mare onoare și recunoscută ca o disciplină fundamentală a științei moderne.

De aceea, ținînd seamă și de faptul că este prima carte de statistică matematică în limba romînă, ea a păstrat caracterul originar de lecții de introducere sistematică și aprofundată a noțiunilor, a principalelor structuri pe care le folosește această știință, a metodelor sale proprii și a principalelor probleme cărora ea are să le răspundă.

O noutate în prezentarea materiei acestei cărți constă în faptul de a fi renunțat la introducerea probabilistică, care era necesară pînă acum și în cărțile din țările cu cunoștințe mai răspîndite de teoria probabilităților. Presupunem cunoscute aceste cunoștințe și pentru aceasta trimitem la cărțile apărute recent în limba romînă. O singură excepție am făcut la Cap. XX, în care prezentînd o teorie de estimație pentru variabilele dependente am expus cu oarecare detalii, teoria proceselor stochastice (lanțuri cu legături complete și Markov) pe care le avem în vedere.

Scanned with OKEN Scanner

In acest volum consacrat teoriei nu am dat decît puține aplicații. In genere, aplicațiile ce s-ar fi putut da ar fi fost simple exerciții numerice, de folos redus. Adevăratele aplicații, utile, trebuie încadrate în domenii experimentale precis conturate, prezentate riguros științific și critic.

Ne propunem să dăm un nou volum, suplimentar acestuia, destinat aplicațiilor în care toate teoriile aici prezentate să-și arate utilitatea, fiecare cu avantajele și dezavantajele sale.

O. Onicescu și Gh. Mihoc

Scanned with OKEN Scanner

TABLA DE MATERII

Pag.

Prefață 3

PARTEA ÎNTÎI

Cap. I. Distribuții statistice

§ 1. Funcție. Funcție empirică 11
§ 2. Distribuții 13
§ 3. Distribuții cu o dimensiune 16
§ 4. Reprezentarea grafică a distribuțiilor cu o dimensiune . 20
§ 5. Curba de distribuție 24
§ 6. Funcția de distribuție 27
§ 7. Cazul general 29

Cap. II. Momente

§ 1. Definiția momentelor și valorilor medii 30
§ 2. Inegalități între valori medii și momente 35
§ 3. Momente centrate 40
§ 4. Funcție caracteristică 44
§ 5. Cumulanți 49
§ 6. Momente factoriale 51
§ 7. Corecțiile momentelor datorită grupării 54

Cap. III. Valori tipice uzuale

§ 1. Considerații generale 62
§ 2. Media aritmetică 62
§ 3. Mediana 64
§ 4. Quantile 69
§ 5. Modul 70
§ 6. Extinderea sau intervalul de variație 72
§ 7. Media abaterilor absolute 73
§ 8. Dispersia 76
§ 9. Media diferențelor absolute 81
§ 10. Coeficienți de variabilitate și concentrație 82

Cap. IV. Distribuții cu două dimensiuni

§ 1. Considerații generale 85
§ 2. Funcția de distribuție 86
§ 3. Distribuții discrete 88
§ 4. Valori medii condiționate 94

Scanned with OKEN Scanner

§ 5. Curbe de regresiune parabolică 96
§ 6. Coeficientul de contingenţă a lui Pearson 110
§ 7. Coeficientul maxim de corelaţie 113
§ 8. Calculul corelaţiilor cu ajutorul dezvoltării după polinoame ortogonale 125

Cap. V. Distribuţii cu mai multe dimensiuni

§ 1. Definirea distribuţiilor 132
§ 2. Distribuţii marginale. Distribuţii condiţionate 134
§ 3. Momente 136
§ 4. Regresiune. Corelaţie parţială 138
§ 5. Rangul unei distribuţii 140
§ 6. Funcţii de regresiune 142
§ 7. Coeficientul multiplu de corelaţie 147
§ 8. Studiul regresiunii cu ajutorul regulii celor mai mici pătrate 149

Cap. VI. Distribuţia normală

§ 1. Consideraţii generale 152
§ 2. Momente 155
§ 3. Funcţia caracteristică 157
§ 4. Polinoamele lui Hermite 158
§ 5. Distibuţia normală multidimensională 159
§ 6. Funcţia caracteristică a distribuţiei normale multidimensionale 162
§ 7. Momentele distribuţiei normale multidimensionale . . . 167

Cap. VII. Distribuţia binomială

§ 1. Definiţii 170
§ 2. Probabilitatea $P_{n,\alpha}$ 172
§ 3. Variaţia probabilităţilor $P_{n,\alpha}$ 175
§ 4. Funcţia de distribuţie 177
§ 5. Momente 178
§ 6. Funcţia caracteristică 181
§ 7. Expresia asimptotică a probabilităţii $P_{n,\alpha}$ 185
§ 8. Distribuţia limită 188
§ 9. O altă demonstraţie a teoremei de limită 192

Cap. VIII. Distribuţia multinomială

§ 1. Consideraţii generale 194
§ 2. Determinarea probabilităţii ca în n probe succesive să obţinem vectorii e_i respectiv de α_i ori 196
§ 3. Valoarea asimptotică a probabilităţii $P_n(\alpha_1, \alpha_2, \ldots, \alpha_m)$. 198
§ 4. Momente 200
§ 5. Momente centrate 204
§ 6. Funcţia caracteristică 209
§ 7. Distibuţia limită 211
§ 8. Funcţia de distribuţie 213

Scanned with OKEN Scanner

Cap. IX. Distribuţia hipergeometrică

§ 1. Consideraţii generale 218
§ 2. Determinarea probabilităţii ca în n probe succesive să obţinem α bile albe şi β bile negre 220
§ 3. Momente 221
§ 4. Momente centrate 226
§ 5. Momente factoriale 227
§ 6. Selecţia dintr-o colectivitate finită 229
§ 7. Valoarea asimptotică a probabilităţii $P_n(\alpha)$ 232
§ 8. Distribuţia hipergeometrică multidimensională 237
§ 9. Momentele distribuţiei hipergeometrice multidimensionale . 242

Cap. X. Distribuţia Poisson, distribuţia binomială cu exponent negativ

§ 1. Consideraţii generale 246
§ 2. Momente 249
§ 3. Funcţia caracteristică 250
§ 4. Distribuţia Poisson multidimensională 252
§ 5. Funcţia caracteristică a distribuţiei Poisson multidimensionale 254
§ 6. O altă generalizare a distribuţiei Poisson 256
§ 7. Momentele distribuţiei Poisson multidimensionale . . 262
§ 8. Distribuţia binomială generalizată a lui Poisson . . . 264
§ 9. Dispersia subnormală 266
§ 10. Dispersia supranormală 269
§ 11. Distribuţia binomială cu exponent negativ 272
§ 12. Momentele distribuţiei binomiale cu exponent negativ . 273
§ 13. Funcţia caracteristică a distribuţiei binomiale cu exponent negativ 275
§ 14. Interpretarea geometrică a distribuţiei binomiale şi a distribuţiei binomiale cu exponent negativ 277

Cap. XI. Alte distribuţii clasice

§ 1. Curbele de distribuţie ale lui Pearson 279
§ 2. Tipuri de curbe Pearson 283
§ 3. Distribuţia χ^2 a lui Helmert-Pearson 286
§ 4. O proprietate de limită a distibuţiei multinomiale . . 294
§ 5. Distribuţii date de forme pătratice de variabile aleatorii . 299
§ 6. Reprezentarea grafică a distribuţiei $H(s, \sigma)$ 302
§ 7. Momentele distribuţiei $H(s, \sigma)$ 305
§ 8. Distribuţii înrudite cu distribuţia $H(s, \sigma)$ 307
§ 9. Distribuţia Student 309
§ 10. Momentele distribuţiei Student 314
§ 11. Alte variabile care dau naştere la distribuţia Student sau la distribuţii înrudite cu ea 316
§ 12. Distribuţia Snedecor 324
§ 13. Distribuţia Beta 332
§ 14. Distribuţia rectangulară 334
§ 15. Distribuţia Cauchy 341

Scanned with OKEN Scanner

§ 16. Distribuţia Wishart 342
§ 17. Distribuţia lui Hotelling 347
§ 18. Teorema lui Rao 351

PARTEA A DOUA

Cap. XII. Teoria selecţiei

§ 1. Principiul metodei selecţiei 354
§ 2. Estimaţii corecte şi absolut corecte 356
§ 3. Media de selecţie 359
§ 4. Momente de selecţie 366
§ 5. Momente centrate de selecţie 372
§ 6. Momente de selecţie pentru o distribuţie multidimensională 380

Cap. XIII. Selecţia dintr-o colectivitate normală

§ 1. Consideraţii generale 385
§ 2. Distribuţia momentului centrat de selecţie de ordinul al doilea 386
§ 3. Distribuţia raportului dintre abaterea mediei de selecţie şi valoarea medie pătratică de selecţie 395
§ 4. Cazul unei distribuţii normale multidimensionale . . . 396

Cap. XIV. Sondaje cu o fază

§ 1. Consideraţii generale 406
§ 2. Sondaje pentru o distribuţie infinită 408
§ 3. Sondaje dintr-o colectivitate finită 420
§ 4. Sondajul simplu 423
§ 5. Sondajul stratificat simplu 437

Cap. XV. Sondaje cu două faze

§ 1. Sondajul cu două faze cu o singură caracteristică . . 455
§ 2. Sondajul cu două faze cu mai multe caracteristici . . 469

PARTEA A TREIA

Cap. XVI. Teoria estimaţiei pentru familiile de distribuţii cu un parametru

§ 1. Problema estimaţiei parametrilor 474
§ 2. Aplicarea teoremei lui Bayes 475
§ 3. Metode directe de estimaţie 483
§ 4. Funcţii de estimaţie eficiente 486
§ 5. Funcţii de estimaţie suficiente 494
§ 6. Statistici exhaustive 499
§ 7. Metoda verosimilităţii maxime 504
§ 8. Metoda momentelor 509
§ 9. Cazul discontinuu 511

Scanned with OKEN Scanner

Cap. XVII. Teoria estimației pentru familiile de distribuții cu mai mulți parametri

§ 1. Considerații generale. Definiții 516
§ 2. Determinarea unei statistici necesare și suficiente . . 518
§ 3. Cazul familiilor de distribuții unidimensionale . . . 519
§ 4. Determinarea densității de distribuție în cazul unei statistici necesare și suficiente de rang finit 521
§ 5. Metoda verosimilității maxime 523
§ 6. Funcții de estimație eficiente 524

Cap. XVIII. Teoria estimației pentru familiile de distribuții frecvent întîlnite în aplicații

§ 1. Schema lui Bernoulli 530
§ 2. Distribuția multinomială 532
§ 3. Distribuția Poisson 534
§ 4. Legea normală. Estimația mediei 535
§ 5. Legea normală. Estimația valorii medii pătratice . . . 539
§ 6. Legea normală. Estimația dispersiei 545
§ 7. Legea normală. Mediana ca funcție de estimație . . . 548
§ 8. Legea normală. Estimația valorii medii și a valorii medii pătratice 550
§ 9. Distribuția lui Cauchy 555
§ 10. Distribuția normală cu două dimensiuni 556

Cap. XIX. Regiuni de încredere

§ 1. Considerații generale 558
§ 2. Diferite aplicații 564
§ 3. Construirea intervalelor de încredere pentru selecții mari . 573

Cap. XX. Teoria estimației pentru variabile dependente

§ 1. Considerații generale 578
§ 2. Lanțuri cu legături complete 579
§ 3. Diferite categorii de lanțuri 583
§ 4. Legea redusă limită a sumelor de variabile înlănțuite în sensul lui Markov 585
§ 5. Calculul asimptotic al funcțiilor caracteristice . . . 587
§ 6. Momentul de ordinul întîi 592
§ 7. Valoarea asimptotică a momentului de ordinul al doilea . 594
§ 8. Comportarea simptotică a funcțiilor caracteristice . . . 595
§ 9. Cazul lanțului Markov finit 595
§ 10. Rădăcinile ecuației caracteristice 599
§ 11. Matrice decompozabile și indecompozabile 605
§ 12. Momentul de ordinul întîi $M_k^{(1)}(n; \theta)$ 607
§ 13. Momentul de ordinul al doilea $M_k^{(2)}(n; \theta)$ 609
§ 14. Legea limită 612
§ 15. Legile limită pentru sumele de vectori înlănțuiți . . . 613
§ 16. Funcții de estimație eficiente pentru lanțuri Markov . . 621

Scanned with OKEN Scanner

§ 17. Cazul unui lanț Markov cu două stări, cu un singur parametru de estimat 631
§ 18. Cazul unui lanț Markov cu mai mulți parametri . . . 636
§ 19. Metoda verosimilității maxime 646
§ 20. Cazul lanțurilor cu legături complete 652
§ 21. Cazul studiat de F. Chartier 654

Cap. XXI. **Procedee de estimare a funcției de distribuție**

§ 1. Teoreme de convergență a frecvenței către probabilitate . 655
§ 2. Convergența funcției $F_n(x)$ către $F(x)$ 662
§ 3. Cîteva teoreme generale de limită 664
§ 4. Tipuri de convergență a funcției de frecvență . . . 674
§ 5. Compararea a două funcții de frecvență 686

Cap. XXII. **Statistica ordinei**

§ 1. Statistică de ordine pentru variabilele exponențiale și omogene 694
§ 2. Aplicațiile la statistica de ordine 701
§ 3. Cîteva teoreme relative la șirul ordonat $Y_1, Y_2, \ldots, Y_n$ în enunțul și demonstrarea lui Șerban Gheorghiu . . . 706

Cap. XXIII. **Verificarea ipotezelor**

§ 1. Considerații generale 710
§ 2. Modul de verificare al ipotezelor 711
§ 3. Erori de genul întîi și al doilea 713
§ 4. Puterea unui test 715
§ 5. Examinarea cazului unui parametru unic care ia două valori 719

Bibliografie 724

Indice alfabetic 725

Scanned with OKEN Scanner

PARTEA INTII

Capitolul I

Distribuții statistice

§ 1. Funcție. Funcție empirică

Fiind date două mulțimi de obiecte a și A se poate caracteriza între ele o dependență în modul următor.

Să numim cu p obiectul generic sau, cum vom mai spune, punctul generic din mulțimea a. Presupunem că aceste puncte pot fi individualizate într-un mod oarecare astfel ca să aibă sens a vorbi despre un anume punct p sau despre un anume punct P. Dacă la fiecare punct p din a corespunde un anumit punct determinat P din A, corespondența aceasta $p \to P$ este o *funcție*. Se scrie $P=f(p)$.

Să considerăm, de pildă, un corp care cade; el face să se stabilească o corespondență între diferitele momente ale timpului $t\ (\equiv p)$ și pozițiile P pe care le ocupă în spațiu. Corespondența $P=f(t)$ este o funcție și reprezintă o lege naturală. Dacă avem de-a face cu o cădere pe verticală, fiecare poziție P poate fi caracterizată printr-un singur număr real, după cum momentele t sînt de asemenea caracterizate prin numere reale. Corespondența este o funcție reală. Dacă avem de-a face cu un proiectil a cărui mișcare este curbă, fiecare poziție este un vector în spațiul cu trei dimensiuni, funcția $P=f(t)$ este vectorială.

Caracterizarea de mai sus este matematică. Faptul că unele legi naturale pot fi exprimate ca funcții determinate $P=f(p)$ nu face decît să izoleze din mulțimea tuturor funcțiilor pe una dintre ele. Aceasta nu trebuie să altereze caracterul pur matematic al corespondenței reprezentate de funcție, în care p are întreaga generalitate a punctelor din mulțimea a și P se consideră ca determinat o dată cu p. Dacă scriem, de pildă, pentru legea căderii corpurilor în vid

$$z=-\frac{1}{2}gt^2+\left(\frac{dz}{dt}\right)_{t=0}t+z_0,$$

această funcţie specială este expresia unei legi naturale după care corpurile în cădere liberă au acceleraţie constantă.

Verificările experimentale ne dau funcţii empirice care fac să corespundă, în cazul de faţă, unei succesiuni de momente înregistrate de un cronometru, să zicem de la 1 la 10 secunde, înălţimile efective

$$z_0, z_1, \ldots, z_{10}$$

la care s-a găsit un corp în cădere. Corespondenţa între numerele efective $1, 2, 3, \ldots, 10$ şi numerele $z_1, z_2, \ldots, z_{10}$ este de asemenea o funcţie, pe care o numim, după originea ei, *empirică*.

Corespondenţa dată de o tabelă de logaritmi între numerele întregi de la 1 la 1 000 000 şi valorile numerice, numite logaritmi în baza 10, constituie o funcţie numerică, care fără să aibă caracterul empiric semnalat mai sus, este numai o corespondenţă între numere speciale, avînd o strînsă legătură cu funcţia teoretică $X = \log x$ care are semnificaţie pentru orice x şi care nu coincide pentru un întreg oarecare ce se găseşte în tabele decît cu o anume aproximaţie cu valoarea dată de aceste tabele.

In general, o experienţă ne dă o funcţie empirică, adică o corespondenţă între un număr finit de elemente $p_1, p_2, \ldots, p_n$ şi elementele $P_1, P_2, \ldots, P_n$ caracterizate fiecare cu eroarea de măsură sau de numărare indisolubil legate de operaţiile practice. Comparaţia acestei funcţii numerice cu funcţia teoretică, funcţia teoretică cunoscută sau numai ipotetic cunoscută, este o cercetare care aparţine uneori statisticii teoretice, dar mai cu seamă celei practice.

In afară de funcţiile de punct $P = f(p)$, un rol deosebit de important în statistică îl joacă *funcţiile de mulţime*.

Lungimea unui segment, aria unui poligon, volumul unui tetraedru sînt funcţii de mulţimea de puncte la care se referă.

Probabilitatea unui eveniment (care constă în general într-o mulţime de evenimente elementare) este o funcţie de mulţime.

Aspectul funcţiilor empirice de mulţime este foarte variat, după felul mulţimilor considerate.

Foarte adesea aceste funcţii empirice sînt *frecvenţe*. Aşa de pildă, numărul picăturilor de ploaie care cad pe o anumită porţiune de suprafaţă este o frecvenţă şi poate da o măsură a acestei suprafeţe.

In particular *funcţia de repartiţie* sau *de distribuţie* a unei variabile aleatorii prezintă un deosebit interes teoretic şi practic.

Scanned with OKEN Scanner

§ 2. Distribuţii

In cele ce urmează vom examina diferite aspecte empirice sau teoretice ale acestei noţiuni de funcţie de distribuţie sau mai scurt *distribuţie.*

Iată cîteva exemple de distribuţii empirice (tabelele I.2,1—I.2,5).

TABELA I.2,1

Distribuţia unui număr de 53 680 familii, avînd fiecare cîte 8 copii, în raport cu numărul băieţilor

Numărul băieţilor x	Numărul familiilor $F(x)$
0	215
1	1 485
2	5 331
3	10 649
4	14 959
5	11 929
6	6 678
7	2 092
8	342
Total	53 680

Extras din R. A. Fischer, Les méthodes statistiques adaptées à la recherche scientifique, p. 61.

TABELA I.2,2

Distribuţia unui număr de 2 000 serii a cîte 500 numere întregi consecutive cuprinse între numerele 9 000 000 şi 10 000 000, în raport cu numărul numerelor prime

Numărul numerelor prime x	Numărul seriilor $F(x)$	Numărul numerelor prime x	Numărul seriilor $F(x)$	Numărul numerelor prime x	Numărul seriilor $F(x)$
18	1	27	130	36	89
19	1	28	135	37	72
20	7	29	171	38	54
21	9	30	177	39	27
22	13	31	210	40	20
23	20	32	194	41	11
24	46	33	171	42	4
25	73	34	143	43	4
26	105	35	111	44	2
				Total	2 000

Extras din R. de M. de Ballore, Probabilités et Statistiques, p. 158.

TABELA I.2,3

Distribuţia precipitaţiilor atmosferice la noi în ţară, în intervalul 1896—1938, în timpul verii, în raport cu cantitatea de apă

Cantitatea de apă, în mm x	$F(x)$	Cantitatea de apă, în mm x	$F(x)$
100—130	2	220—250	6
130—160	6	250—280	1
160—190	6	280—310	3
190—220	15	310—340	1
		Total	40

Extras din Anuarul Statistic al Rominiei, 1939.

Scanned with OKEN Scanner

TABELA I.2,4

Distribuţia planetelor în raport cu mărimea paralaxei

Intervalele valorilor paralaxei p (cu unitatea 0″,001) x	Numărul planetelor observate $F(x)$	Intervalele valorilor paralaxei p (cu unitatea 0″,001) x	Numărul planetelor observate $F(x)$	Intervalele valorilor paralaxei p (cu unitatea 0″,001) x	Numărul planetelor observate $F(x)$
— 30 la — 25	3	10 la 15	56	55 la 60	12
— 25 la — 20	7	15 la 20	50	60 la 65	17
— 20 la — 15	7	20 la 25	43	65 la 70	13
— 15 la — 10	17	25 la 30	45	70 la 75	15
— 10 la — 5	27	30 la 35	43	75 la 80	12
— 5 la 0	39	35 la 40	32	80 la 85	6
0 la 5	46	40 la 45	33	85 la 90	6
5 la 10	51	45 la 50	29	90 la 95	5
		50 la 55	21	95 la 100	5

Extras după Stellar Dynamics de W. M. Smart.

TABELA I.2,5

Distribuţia populaţiei urbane la noi în ţară, în raport cu vîrsta

Vîrsta, ani x	Numărul locuitorilor $F(x)$	%	Vîrsta, ani x	Numărul locuitorilor $F(x)$	%	Vîrsta, ani x	Numărul locuitorilor $F(x)$	%
0— 4	351 004	9,6	20—24	464 457	12,7	40—49	385 565	10,7
5— 9	335 794	9,2	25—29	356 697	9,8	50—64	331 370	9,1
10—14	270 149	7,4	30—34	262 428	7,2	65 în	179 129	4,8
15—19	449 573	12,3	35—39	261 933	7,2	sus	Total	100,0

Extras din Recensămîntul general al populaţiei din Romînia din anul 1930.

Din aceste exemple se observă că o distribuţie se referă la o *colectivitate*, care reprezintă o mulţime de elemente avînd unele caracteristici comune. Caracteristicile comune pe baza cărora colectivitatea este împărţită în grupe sau clase se numesc *variabile* sau *argumente*. Numărul elementelor unei grupe sau clase poartă numele de *frecvenţă absolută* a acelei grupe sau clase, înăuntrul colectivităţii, iar numărul tuturor elementelor colectivităţii, poartă numele de *volum* al ei.

Scanned with OKEN Scanner

Astfel în tabela I. 2, 1 variabila sau argumentul considerat este numărul băieţilor (0, 1, 2, . . . , 8), frecvenţele absolute sînt date de numărul familiilor care au acelaşi număr de copii (215, 1 485, ... , 342); volumul este numărul total de 53 680 familii observate. Tot astfel în tabela I.2,3, argumentul este dat de cantitatea de apă măsurată în intervale egale de cîte 30 mm (100—130, 130—160, ... , 310—340), frecvenţele absolute sînt date de numărul precipitaţiilor atmosferice corespunzătoare intervalelor arătate (2, 6, 6, 15 ,..., 1), iar volumul este numărul total de 40 precipitaţii atmosferice.

Există deci o corespondenţă între valorile argumentului şi valorile frecvenţelor respective: la fiecare valoare a argumentului corespunde o valoare bine determinată a frecvenţei. Frecvenţa este deci o funcţie, determinată empiric sau teoretic cum vom vedea în numeroase împrejurări, specific statistică.

Să observăm că în tabelele I.2,1 şi I.2,2 pentru fiecare valoare x este dată valoarea corespunzătoare a lui $F(x)$, pe cînd în tabelele I.2,3, I.2,4 şi I.2,5 $F(x)$ reprezintă valoarea corespunzătoare unor intervale. In acest caz, în distribuţiile considerate, $F(x)$ reprezintă în realitate o funcţie de aceste intervale.

In unele cazuri, pentru a ne face o idee mai precisă de importanţa clasei corespunzătoare faţă de mulţimea de elemente observate, frecvenţele absolute se raportează la totalul lor. Se obţin, în modul acesta, ceea ce se cunoaşte sub numele de *frecvenţe relative* sau, mai scurt, frecvenţe. Calculul frecvenţelor relative se face de obicei prin proporţia de 1, 10, 100 etc. Notînd cu $F(x)$ frecvenţa absolută a unei clase, cu $\sum F(x)$ suma frecvenţelor (volumul colectivităţii), frecvenţa relativă este dată de formula

$$f = p\frac{F(x)}{\sum F(x)} \qquad (p = 1, 10, 100, \ldots,).$$

Un exemplu de frecvenţe relative ni-l oferă tabela I.2,5 în coloana a 3-a, unde aceste frecvenţe sînt calculate sub formă de procente (9,6; 9,2; ...; 4,8).

Este clar că valoarea frecvenţei relative poate pune mai bine în evidenţă semnificaţia cantitativă a clasei corespunzătoare.

O privire generală asupra argumentelor distribuţiilor date mai sus ne arată că există o deosebire importantă între aceste argumente. Aşa, de exemplu, în tabelele I.2,1 şi I.2,2, argumentul ia numai valori întregi (de la 0 la 8 în prima tabelă şi între 18 şi 44 în tabela a doua), în timp ce în tabelele I.2,3, I.2,4 şi I.2,5 argumentul poate

Scanned with OKEN Scanner

lua orice valoare cuprinsă în scara lui de variaţie (între 100 şi 340 în tabela I.2,3, respectiv între —30 şi 100 în tabela I.2,4, între 0 şi 65 în tabela I.2,5).

În prima categorie argumentul este *discontinuu*, iar în categoria a doua argumentul este *continuu*. În practică, continuitatea argumentului depinde de precizia cu care se pot efectua măsurătorile. Astfel, într-o colectivitate de indivizi talia poate lua principial orice valoare, dar practic ea se măsoară în centimetri sau, cel mult, în milimetri.

După felul argumentului (discontinuu sau continuu) deosebim *distribuţii discrete* şi *distribuţii continue*. Distribuţiile arătate în tabelele I.2.1 şi I.2,2 sînt distribuţii discrete, pe cîtă vreme cele din tabelele I.2,3, I.2,4 şi I.2,5 sînt continue.

Clasificarea unei distribuţii după argumente poate fi făcută şi din punct de vedere *calitativ* sau *cantitativ*. În primul caz, distribuţia poartă numele de *homogradă*, iar în al doilea caz se numeşte *heterogradă*.

Exemplele considerate pînă acum sînt distribuţii heterograde, deoarece argumentul este exprimat cantitativ prin numere.

În vederea studiului distribuţiilor este necesară clasificarea lor în raport cu numărul variabilelor. Astfel avem distribuţii de o singură variabilă sau *unidimensionale*, distribuţii *bidimensionale* şi,în general, distribuţii *multidimensionale*.

§ 3. Distribuţii cu o dimensiune

Fie X o distribuţie relativă la o variabilă x, care poate lua valorile $x_1, x_2, \ldots, x_n$ şi $s_1, s_2, \ldots, s_n$ frecvenţele absolute corespunzătoare. Schematic vom indica această distribuţie în modul următor:

$$X\begin{pmatrix} x_1, x_2, \ldots, x_n \\ s_1, s_2, \ldots, s_n \end{pmatrix}$$

sau, mai simplu,

$$X\begin{pmatrix} x_i \\ s_i \end{pmatrix} \qquad (i=1,2,\ldots,n).$$

Dacă notăm cu S volumul şi cu f_i $(i=1,2,\ldots,n)$ frecvenţele relative ale distribuţiei X, avem

$$S=\sum_{i=1}^{n} s_i; \quad f_i=\frac{s_i}{S}=\frac{s_i}{\sum_{i=1}^{n} s_i} \qquad (i=1,2,\ldots,n).$$

Scanned with OKEN Scanner

Distribuţiile întîlnite în paragraful precedent sînt toate cu o dimensiune sau de o singură variabilă (argument) care, aşa cum s-a arătat, pot fi discontinue, ca cele din tabelele I.2,1 şi I.2,2 sau continue, ca acelea din tabelele I.2,3, I.2,4 şi I.2,5.

Valorile variabilei sînt, în general, cuprinse în limitele unui interval dat (α, β), care se numeşte *oscilaţia distribuţiei*. De exemplu, valorile variabilei care intervine în precipitaţiile atmosferice (tabela I.2,3) sînt cuprinse între limitele 100 şi 340 mm. Fiind vorba de o distribuţie principial continuă, practic limitată la un număr de mm, variabila poate, evident, lua orice valoare între aceste două limite. Dacă am fi trecut în tabelă toate valorile variabilei măsurate în milimetri, am fi obţinut o tabelă cu 240 clase în loc de opt. O astfel de tabelă este prea mare, dacă nu vrem să semnalăm şi faptul că scoate în evidenţă prea multe neregularităţi ale colectivităţii.

Pentru remedierea acestor neajunsuri, se împarte intervalul (α, β) într-un număr suficient de intervale mai mici.

$$(\alpha, x_1), \quad (x_1, x_2), \ldots, (x_{i-1}, x_i), \ldots, (x_n, \beta)$$

în raport cu valorile x_i $(i = 1, 2, \ldots, n)$ ale variabilei, care verifică inegalităţile $\alpha < x_i < \beta$. Aşa s-a alcătuit tabela I.2,3, unde intervalul (100, 340) s-a împărţit în opt intervale mai mici, în raport cu numerele 130, 160, ..., 310.

In ceea ce priveşte numărul intervalelor de clasă ale unei distribuţii de o variabilă, acesta depinde de caracterele specifice ale distribuţiei respective. In toate cazurile, numărul acestor intervale trebuie fixat în aşa fel ca să punem cît mai în evidenţă caracterul regulat al distribuţiei. Aceasta fiind o chestiune de tehnică a cercetărilor statistice, nu insistăm aici mai mult asupra ei.

Important pentru studiul distribuţiilor de o variabilă este modul de alcătuire al intervalelor şi mărimea lor, deoarece de felul cum este fixată scara unei distribuţii depinde punerea în evidenţă a particularităţilor ei. Pentru multe distribuţii intervalele trebuie să fie egale pentru a putea face o comparaţie cît mai bună între frecvenţele claselor. Exemple sînt date de distribuţia precipitaţiilor atmosferice (tabela I,2,3) şi distribuţia planetelor (tabela I.2,4), unde intervalele sînt egale cu cîte 30 mm, respectiv cu 0″,005. Pentru distribuţia populaţiei, însă (tabela I.2,5), intervalele sînt inegale şi anume, primele opt intervale sînt egale, al nouălea este de două ori mai mare decît fiecare dintre primele opt, iar al zecelea de trei ori mai

Scanned with OKEN Scanner

mare decît fiecare dintre acestea. In sfîrşit, ultimul interval este deschis.

In cazuri ca acesta al distribuţiei populaţiei după vîrstă renunţăm la intervale egale, deoarece folosirea intervalelor de cîte 5 ani pentru toată scara distribuţiei ar da naştere unei tabele prea lungi, fără a reda esenţialul, în special pentru studiul distribuţiei populaţiei urbane după virsta de 40 ani, cînd frecvenţele variază mai repede.

Există colectivităţi pentru care intervalele claselor marginale sînt la rîndul lor împărţite în subintervale egale. Aceasta deoarece elementele care compun clasele marginale au particularităţi speciale, care trebuie subliniate în mod deosebit.

Aşa, de pildă, studiul mortalităţii infantile ne conduce la întocmirea distribuţiilor la vîrstele mici cu subintervale. Exemplul din tabela I.3,1 justifică împărţirea primului interval marginal în alte cinci subintervale egale.

TABELA I.3,1

Distribuţia deceselor de scarlatină în Anglia şi Ţara Galilor în 1933, în raport cu vîrsta

Vîrsta, în ani x	Numărul deceselor $F(x)$		Vîrsta, în ani x	Numărul deceselor $F(x)$
0— 1	16		40 - 45	10
1— 2	69		45—50	6
2— 3	89	322	50—55	7
3— 4	74		55—60	5
4— 5	74		60—65	—
5—10	213		65—70	1
10—15	70		70—75	1
15—20	27		75—80	1
20—25	26		80—	—
25—30	17			
30—35	12		Total	729
35—40	11			

Extras din The Advanced Theory of Statistic, M. G. Kendall, p. 5.

Impărţirea primului interval (interval marginal) în cinci subintervale egale era necesară aici pentru cunoaşterea completă a mortalităţii, deoarece numai în felul acesta se poate observa că mortalitatea nu descreşte uniform în primii cinci ani. De aici rezultă că maximul de decese nu corespunde primului an, cum s-ar fi putut trage con-

Scanned with OKEN Scanner

cluzia dacă, în locul celor cinci subintervale, s-ar fi considerat un singur interval corespunzător totalului de 322 frecvenţe.

De asemenea, important este modul cum trebuie interpretate limitele intervalelor de clasă ale distribuţiei. Se întrebuinţează, de obicei, următoarele forme pentru scrierea intervalelor:

(I)
mai mare sau egal cu 0 şi mai mic decît 1
mai mare sau egal cu 1 şi mai mic decît 2
mai mare sau egal cu 2 şi mai mic decît 3
. . . . . .

(II)
0
0—0,9
1—1,9
2—2,9
. . .

(III)
0—1
1—2
2—3
. .

(IV)
0—
1—
2—
. .

Forma (I) este cea mai corectă, spre deosebire de formele (II), (III) şi (IV) unde se pot ivi ambiguităţi. Cu toate acestea, forma (I) este întîlnită din ce în ce mai rar în distribuţiile statistice, din cauza dezavantajului de a fi greoaie şi nesugestivă.

In general, în distribuţiile de o variabilă continuă se adoptă forma (III) şi (IV), iar în cele de o variabilă discontinuă, forma (II). De altfel, forma (II) este singura care dă o informaţie asupra preciziei în măsurătorile efectuate, ceea ce face ca ea să fie întrebuinţată şi pentru distribuţiile continue, care în practică se prezintă sub forma distribuţiilor discontinue din cauza limitei preciziei cu care se face măsurătoarea.

Se înţelege de la sine că în această ultimă notaţie zecimala 9 nu arată decît că marginea superioară a intervalului considerat nu este atinsă.

Dacă se adoptă una din formele (II), (III) sau (IV), este necesar să se arate în mod precis cui aparţin valorile marginilor fiecărui interval. Aceste valori pot aparţine marginii inferioare, marginii superioare, sau tot aşa de bine pot fi repartizate jumătate intervalului considerat şi jumătate intervalului vecin. Luînd ca exemplu, pentru fixarea ideilor, intervalul (x_i, x_{i+1}), vom atribui acestui interval toate valorile lui x care verifică inegalităţile

$$x_i \leqq x < x_{i+1} \quad \text{(în primul caz)},$$

$$x_i < x \leqq x_{i+1} \quad \text{(în cazul al doilea)},$$

iar în ultimul caz, jumătate din valorile lui x se vor atribui intervalul (x_{i-1}, x_i) şi jumătate intervalului (x_i, x_{i+1}).

Am arătat mai sus importanţa intervalelor marginale pentru studiul distribuţiilor de o variabilă. In multe cazuri, marginile acestor intervale nu sînt fixate. O distribuţie cu asemenea intervale poartă numele de *distribuţie deschisă* la unul sau la ambele capete. Dacă marginile intervalelor marginale sînt fixate, atunci distribuţia se numeşte *închisă*. Primele patru distribuţii prezentate în § 2 sînt închise, iar ultima (a 5-a) este deschisă la un singur capăt.

§ 4. Reprezentarea grafică a distribuţiilor cu o dimensiune

In studiul distribuţiilor cu o dimensiune, reprezentarea grafică este un auxiliar de primul ordin. Intr-adevăr, cu ajutorul reprezentării grafice, se pot observa mai uşor unele proprietăţi ale dependenţei

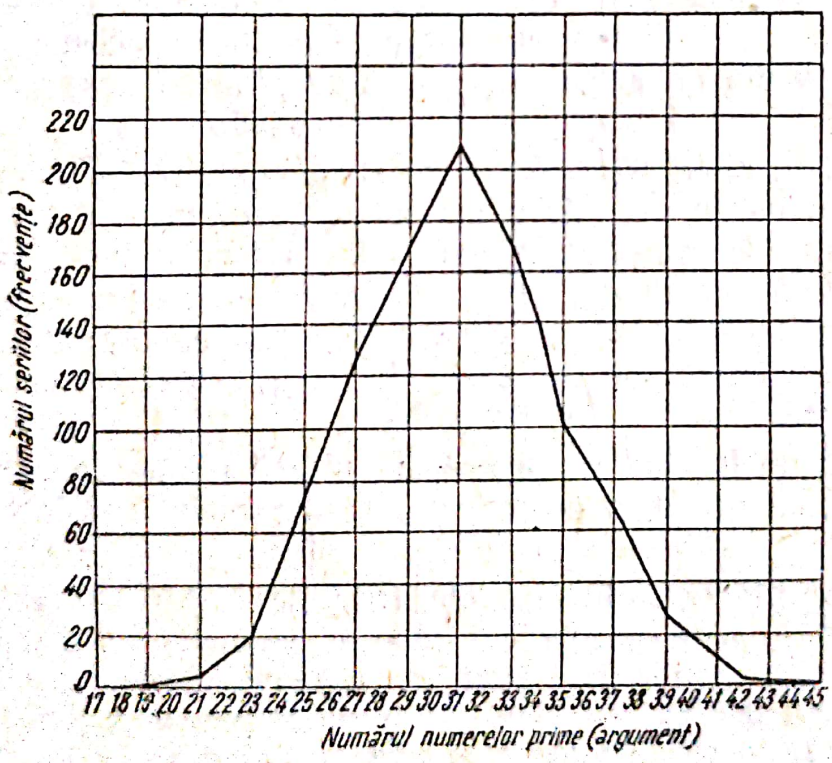


Fig. I.4,1. Poligonul frecvenţelor

funcţionale dintre argumentul şi frecvenţele distribuţiei. Această dependenţă, reliefată şi în distribuţiile prezentate obişnuit sub forma unei tabele, este, în general, bine scoasă în evidenţă prin reprezentarea grafică.

Scanned with OKEN Scanner

In practica statistică se întrebuinţează variate metode de reprezentare grafică a distribuţiil r cu o dimensiune, dar printre cele mai însemnate sînt metodele grafice cunoscute sub numele de *poligonul frecvenţelor* şi *histograma.*

Poligonul frecvenţelor este graficul ce se obţine în sistemul axelor rectangulare, unind printr-o linie poligonală punctele ale căror abcise şi ordonate sînt, respectiv, mulţimea valorilor luate de argument şi frecvenţa corespunzătoare.

Figura I.4,1 reprezintă poligonul frecvenţelor în distribuţia celor 2000 serii cuprinse între numerele 9 000 000 şi 10 000 000 în raport cu numărul numerelor prime (v. tabela I.2,2).

Histograma este reprezentarea grafică rezultată din considerarea frecvenţelor distribuţiei ca arii ale unor dreptunghiuri. De aici rejese că histograma se întrebuinţează pentru reprezentarea grafică a distribuţiilor cu valorile argumentului grupate în intervale de clasă, care se iau ca baze ale dreptunghiurilor. Inălţimile dreptunghiurilor pot fi, evident, egale sau proporţionale cu frecvenţele distribuţiei, după cum bazele dreptunghiurilor se exprimă prin una sau mai multe unităţi de acelaşi fel. De asemenea, se poate folosi şi poligonul frecvenţelor pentru graficul acestor distribuţii. In cazul de faţă, se vor lua ca abscise valorile mijlocului intervalelor, iar ca ordonate valori egale sau proporţionale cu frecvenţele.

Figura I.4,2 este histograma distribuţiei planetelor în raport cu valorile paralaxei (v. tabela I.2,4).

Din cele arătate mai sus se observă că se poate trece uşor de la histograma distribuţiei la poligonul frecvenţelor şi invers. Intr-adevăr, unind prin segmente de dreaptă mijlocul bazelor superioare ale dreptunghiurilor histogramei, se obţine poligonul frecvenţelor. Printr-o operaţie inversă, se poate obţine uşor, din poligonul frecvenţelor, histograma distribuţiei.

Reprezentarea grafică permite, în acelaşi timp, să se răspundă unei probleme importante şi anume aceea relativă la forma distribuţiei.

In cele ce urmează, prezentăm succint cinci tipuri de forme ale unei distribuţii cu o dimensiune, apreciindu-le ca cele mai caracteristice din mulţimea de forme care se întîlnesc în practică.

1. *Distribuţia simetrică* (fig I.4.3.) este distribuţia în care avem aceleaşi valori pentru frecvenţele situate la distanţe egale, de o parte şi de alta a valorii de frecvenţă maximă. De fapt nu există distribuţii care să corespundă riguros acestei definiţii, deoarece, în general, valorile frecvenţelor distribuite în modul arătat faţă de

frecvenţa maximă, diferă, într-o oarecare măsură, unele de altele. Distribuţiile simetrice sînt foarte puţine şi cele mai răspîndite sînt

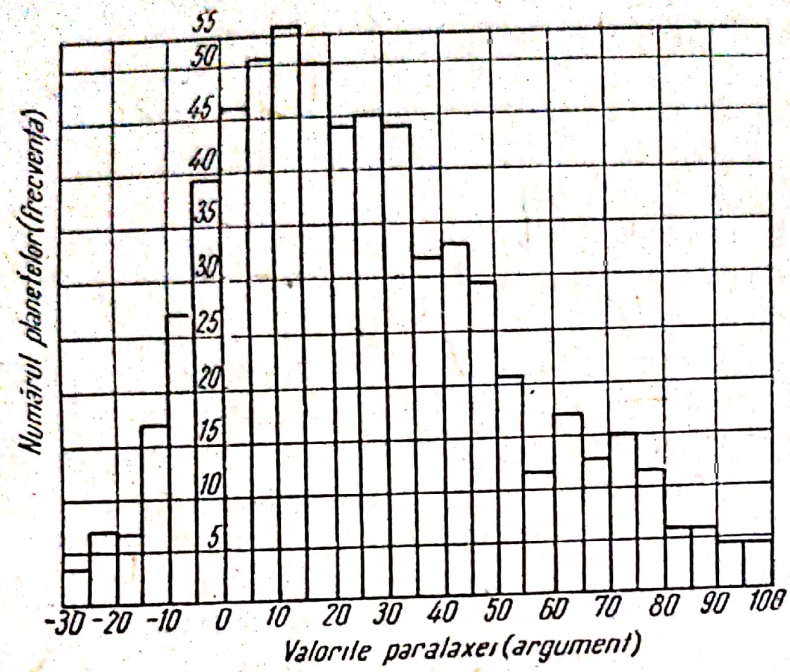


Fig. I.4,2. Histograma

cunoscute sub numele de *distribuţii normale*, studiate prima oară de Laplace şi apoi de Gauss. Astfel de exemple ne oferă biometria, antropometria, precum şi unele măsurători, observaţii etc.

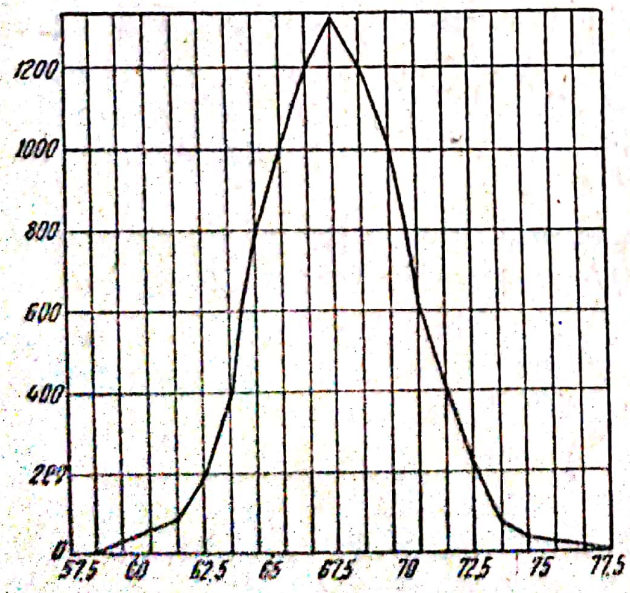


Fig. I.4,3. Distribuţie simetrică

2. *Distribuţia uşor asimetrică* (fig. I.4,1) este aceea în care valorile frecvenţelor situate la distanţe egale de o parte şi de alta a frecvenţei maxime diferă mai mult între ele. Asemenea distribuţii sînt cele mai numeroase şi aproape că nu există domeniu în care să nu fie întîlnite.

3. *Distribuţia pronunţat asimetrică* (fig. I.4,4) este distribuţia în care majoritatea frecvenţelor este concentrată de o singură parte a frecvenţei maxime, care corespunde la o valoare a argumentului

foarte aproape de unul din capetele intervalului de oscilaţie a distribuţiei. Distribuţiile pronunţat asimetrice se prezintă mai rar în statistică.

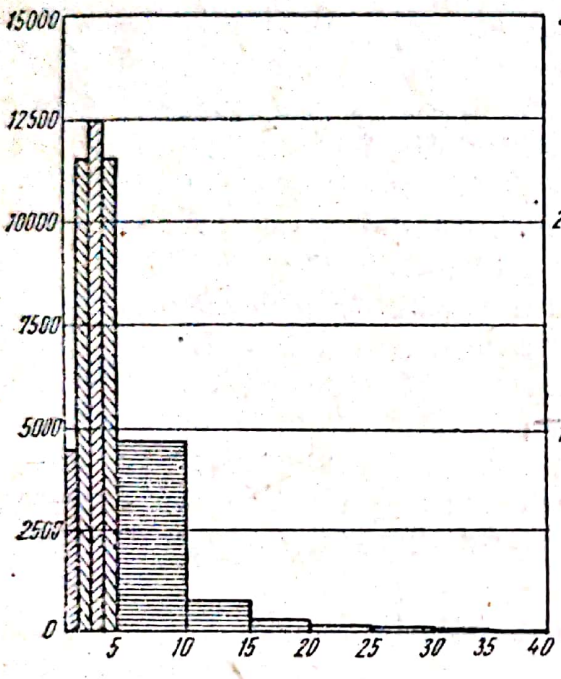


Fig. I.4,4. Distribuţie asimetrică

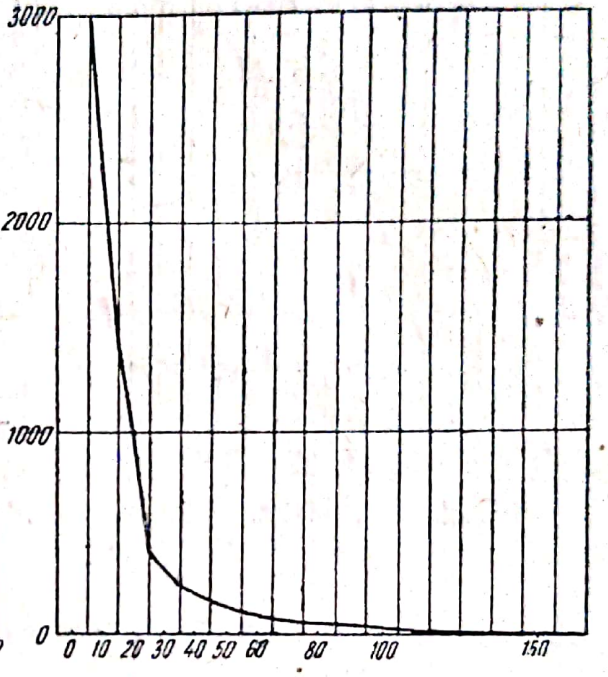


Fig. I.4,5. Distribuţie în *J*

4. *Distribuţia în formă de J* (fig. I.4,5) în care toate frecvenţele sînt concentrate de o singură parte a frecvenţei maxime, corespunzînd uneia din valorile argumentului din extremităţile scării. Aceste distribuţii trebuie cercetate cu multă atenţie, deoarece unele distribuţii pronunţat asimetrice pot apărea ca distribuţii în *J*. Astfel de situaţii se întîmplă de obicei în cazul distribuţiilor continue, din cauza faptului că împărţirea intervalului marginal în subintervale mai mici poate merge oricît de mult. Este de la sine înţeles că toate aceste consideraţii nu mai au obiect în cazul distribuţiilor discrete. Unii autorii nu consideră distribuţiile în *J* ca un tip de formă aparte, ci le înglobează în tipul de formă al distribuţiilor pronunţat asimetrice.

5. *Distribuţia în formă de U* (fig. I.4,6) poate fi simetrică sau asimetrică (uşor sau mai pronunţat) în raport cu o anumită valoare corespunzînd celei mai mici frecvenţe. In aceste distribuţii, frecven-

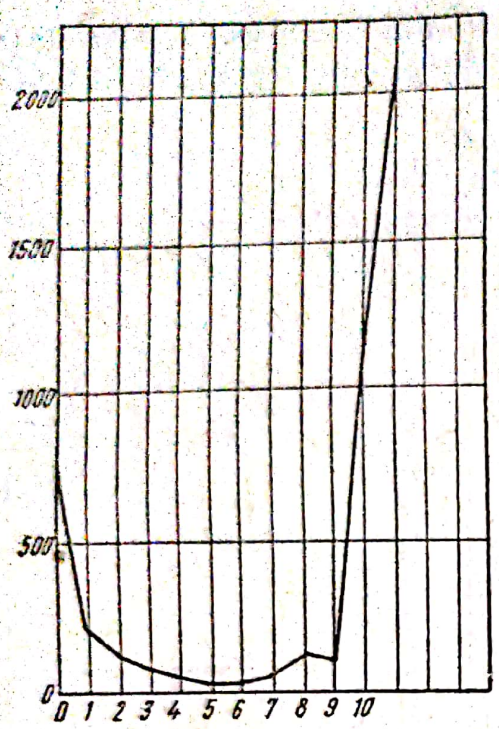


Fig. I.4,6. Distribuţie în U

ţele referitoare la valorile (intervalele) marginale sînt maxime. Distribuţiile în U sînt dintre cele mai rare. Unul dintre fenomenele rare care se manifestă sub forma aceasta este distribuţia numărului de zile după gradul de înnorare a cerului, care a fost pusă în evidenţă întîia oară de către K. Pearson.

6. Deosebit de cele cinci forme de distribuţii mai există şi aşa-zisele *distribuţii cu mai multe vîrfuri*, care se explică prin lipsa de omogenitate a colectivităţii, ca de pildă suprapunerea mai multor specii etc. In fig. I.4,2 avem o distribuţie cu cinci vîrfuri, corespunzătoare intervalelor : 10—15 ; 25—30 ; 40—45 ; 60—65 ; 70—75.

§ 5. Curba de distribuţie

Din expunerea de pînă acum rezultă faptul că cea mai mare parte a distribuţiilor cu o dimensiune au un mare grad de regularitate. In special reprezentarea grafică a acestor distribuţii sub forma poligonului frecvenţelor şi a histogramei ne sugerează ideea de a privi distribuţiile statistice ca prime aproximaţii ale unor distribuţii teoretice, eventual chiar analitice şi reprezentate grafic prin curbe continue sau prin succesiuni de bucăţi de curbe continue. In felul acesta, sîntem conduşi la noţiunea de *curbă de distribuţie* de care ne vom ocupa mai jos.

Să considerăm o variabilă aleatorie x continuă, luînd orice valoare din mulţimea valorilor intervalului finit (α, β), pe care-l împărţim în alte n intervale intermediare egale fiecare cu Δx,

$$(\alpha, x_1), (x_1, x_2), \ldots, (x_{n-1}, \beta),$$

cu ajutorul şirului crescător de valori

$$\alpha < x_1 < x_2 < \ldots < x_{n-1} < \beta.$$

Să presupunem că experienţa care dă naştere variabilei aleatorii x se compune din S probe independente relativ la această variabilă. No-

tînd cu $s_1, s_2, \ldots, s_n$, frecvenţele rezultate din această experienţă obţinem distribuţia

$$X\begin{pmatrix}(\alpha, x_1), & (x_1, x_2), \ldots, & (x_{n-1}, \beta)\\ s_1, & s_2 \quad , \ldots, & s_n\end{pmatrix}.$$

Dacă variabila aleatorie x are o densitate de probabilitate $f(x)$, probabilitatea ca x să ia o valoare cuprinsă în intervalul elementar (x_{i-1}, x_i) este

$$p_i = \int_{x_{i-1}}^{x_i} f(x)\,dx. \tag{I.5,1}$$

In baza legii numerelor mari (teorema lui Bernoulli), putem afirma că

$$P\left(\left|\frac{s_i}{S} - p_i\right| < \varepsilon\right) \text{ este aproape de 1} \tag{I.5,2}$$

pentru ε suficient de mic, dacă numărul probelor este suficient de mare.

Din (I.5,1) şi (1.5,2) rezultă, deci, în condiţii bine cunoscute, relaţia aproximativă

$$\frac{s_i}{S} \simeq \int_{x_{i-1}}^{x_i} f(x)\,dx. \tag{I.5,3}$$

Prezentăm mai jos printr-o histogramă distribuţia X, ţinînd seama că sîntem în cazul unei distribuţii continue cu valorile argumentului grupate în intervalele considerate. Pe acelaşi grafic reprezentăm şi curba a cărei ecuaţie este $y = f(x)$ (fig. I.5,1).

Din relaţia (I.5,3) obţinem, aplicînd teorema mediei

$$\frac{s_i}{S} = f(\xi_i)\Delta x, \quad \text{unde} \quad x_{i-1} \leqslant \xi_i \leqslant x_i. \tag{I.5,4}$$

Această egalitate pune în evidenţă un fapt deosebit de important şi anume că, în anume condiţii de regularitate, histograma este cu atît mai aproape de graficul exact, cu cît numărul S de probe independente este mai mare şi, deci, cu atît mai mică este mărimea co-

mună Δx a fiecărui interval, bineînţeles dacă $f(\xi_i)$ găsit empiric este egal cu valoarea teoretică sau diferă de ea cu o valoare neglijabilă.

Din relaţia (I.5,4) deducem

$$\frac{s_i}{S\Delta x} = f(\xi_i). \tag{I.5,5}$$

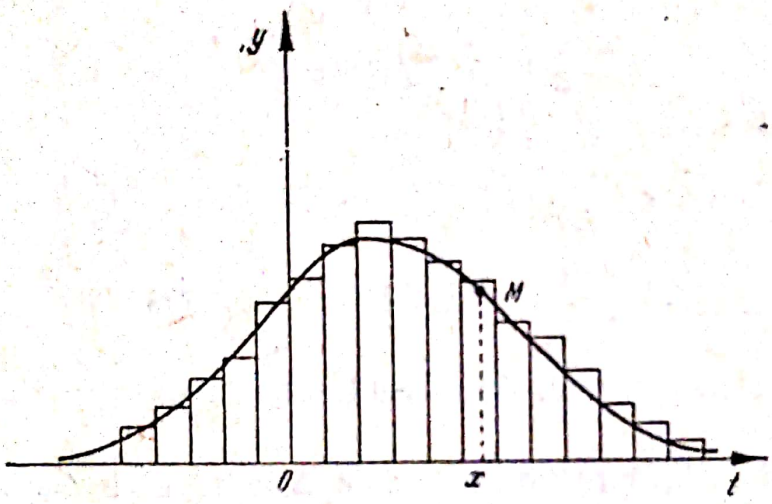


Fig. I.5,1

Înălţimea $\frac{s_i}{S\Delta x}$ a dreptunghiului histogramei, construit pe Δx ca bază şi avînd aria egală cu frecvenţa relativă $\frac{s_i}{S}$, reprezintă cu o aproximaţie care, în condiţii de suficientă regularitate, trebuie să fie de ajuns de bună, ordonata unui punct convenabil ales pe curba $y=f(x)$.

In general, nu cunoaştem, în prealabil, curba de distribuţie şi se cere, plecînd de la datele simple ale înregistrării statistice, determinarea acestei curbe ceea ce constituie una dintre problemele importante ale statisticii matematice.

Cunoaşterea curbei de distribuţie rezolvă problema calculului frecvenţelor probabile, corespunzătoare valorilor variabilei aleatorii x cuprinse între două limite date a şi b $(\alpha<a<b<\beta)$.

Pentru aceasta, să revenim la experienţa care dă naştere variabilei aleatorii x prin cele S probe independente şi să notăm cu $S(a, b)$ frecvenţa referitoare la valorile lui x care aparţin intervalului (a, b).

Urmînd aceleaşi procedee ca şi în cazul frecvenţei s_i pentru valorile lui $x \in \Delta x$, rezultă, de asemenea, relaţia aproximativă

$$\frac{s(a,\ b)}{S} \simeq \int_a^b f(x)\,dx$$

pentru S destul de mare, sau

$$s(a,\ b) = S\int_a^b f(x)\,dx. \qquad (I.5,6)$$

Relaţia (I.5,6) indică probabilitatea de a calcula orice valoare a frecvenţei experimentale pentru valorile argumentului cuprinse între două limite date a şi b, dacă se cunoaşte curba de distribuţie $y=f(x)$ sau histograma pentru un număr mare de probe independente.

§ 6. Funcţia de distribuţie

Să notăm cu $y=f_0(x)$ ecuaţia liniei în scară care reprezintă histograma distribuţiei şi cu $y=f(x)$ ecuaţia curbei, eventuale, teoretice căreia îi corespunde cea de sus.

Presupunem că distribuţia este extinsă în intervalul

$$\alpha < x < \beta.$$

In acest caz

$$F_0(x) = \frac{1}{S_0}\int_\alpha^x f_0(a)\,da$$

este funcţia de distribuţie empirică şi

$$F(x) = \frac{1}{S}\int_\alpha^x f(a)\,da$$

este funcţia de distribuţie teoretică, dacă

$$S_0 = \int_\alpha^\beta f_0(a)\,da; \qquad S = \int_\alpha^\beta f(a)\,da.$$

In cazul distribuţiilor discrete, funcţia de distribuţie pentru valorile variabilei mai mici sau egale cu x_p este

$$F(x_p) = \sum_{i=1}^{p} f(x_i).$$

Ca exemplu, prezentăm graficul funcţiei $F_0(x)$ (fig. I.6,1) pentru distribuţia celor 53 680 familii observate, avînd fiecare 8 copii, în raport cu numărul băieţilor.

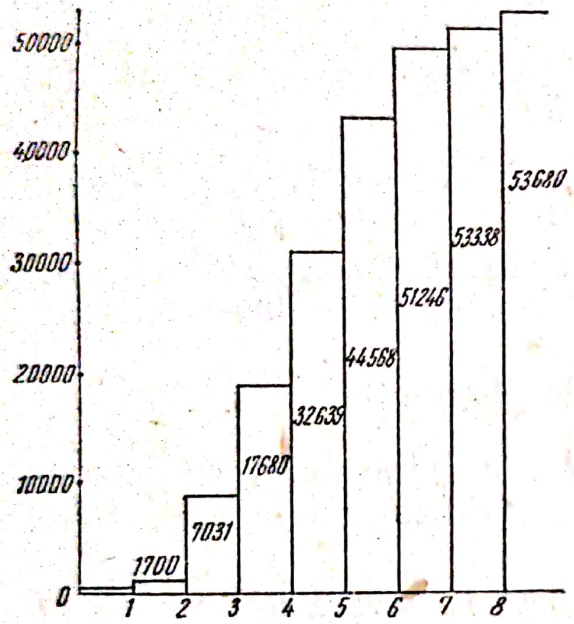


Fig. I.6,1

Nr. băieţilor x	Nr. familiilor s	Frecvenţe cumulate Σs
0	215	215
1	1 485	1 700
2	5 331	7 031
3	10 649	17 680
4	14 959	32 639
5	11 929	44 568
6	6 678	51 246
7	2 092	53 338
8	342	53 680

Din cercetarea acestui grafic, rezultă următoarele valori pentru funcţia de distribuţie:

$$\begin{aligned}
F_0(x) &= 0 \quad \text{dacă} \quad x < 0,\\
F_0(x) &= 215 \quad \text{dacă} \quad 0 \leq x < 1,\\
F_0(x) &= 1\,700 \quad \text{dacă} \quad 1 \leq x < 2,\\
&\cdots\cdots\cdots\\
F_0(x) &= 53\,338 \quad \text{dacă} \quad 7 \leq x < 8,\\
F_0(x) &= 53\,680 \quad \text{dacă} \quad x > 8.
\end{aligned}$$

Distribuţia examinată fiind discretă, funcţia de distribuţie are nouă puncte de discontinuitate (0, 1, 2, . . . , 8) numite *salturi*, din care cauză apare sub forma unei *funcţii cu trepte*.

Scanned with OKEN Scanner

In cazul distribuţiilor continue, graficul funcţiei de distribuţie se prezintă ca o funcţie continuă şi monoton crescătoare în intervalul considerat (fig. I.6,2).

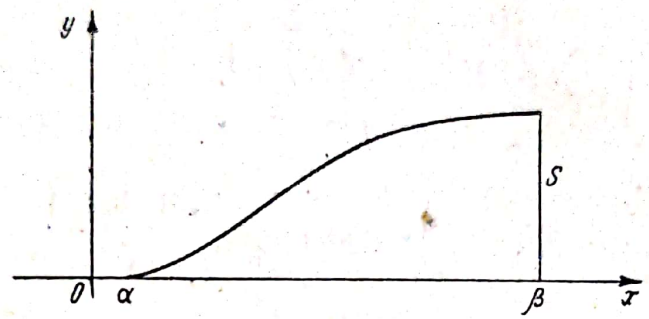


Fig. I.6,2

In expunerea noastră am considerat funcţia de distribuţie continuă la stînga.

§ 7. Cazul general

Progresele realizate în statistica matematică au fost în mare parte posibile datorită faptului că distribuţia empirică a fost pusă în legătură cu funcţia de distribuţie a unei variabile aleatorii. In modul acesta a fost depăşită faza statisticii descriptive, în care numai se descrie fenomenul statistic, dar nu se poate face nici o previziune. A defini o distribuţie statistică matematică înseamnă a cunoaşte legea de probabilitate care stă la baza ei. Din această cauză, putem considera statistica matematică ca o ramură a teoriei probabilităţilor. Totuşi, metodele şi problemele specifice ale statisticii matematice fac din ea o disciplină matematică de sine stătătoare.

Fiind definit pe axa reală R_1 un cîmp de probabilitate şi variabila aleatorie u, funcţia de repartiţie din teoria probabilităţilor este dată de relaţia

$$P(u \leqslant x) = F(x),$$

prin $P(u \leqslant x)$ notînd probabilitatea inegalităţii $u \leqslant x$. De aici înainte, folosind terminologia statisticii matematice, vom numi pe $F(x)$ *funcţie de distribuţie.*

$F(x)$ este o funcţie pozitivă nedescrescătoare, continuă la dreapta, astfel încît $F(-\infty)=0$, $F(+\infty)=1$.

In cazul unei distribuţii continue, derivata funcţiei $F(x)$ există în fiecare punct şi poartă numele de *funcţie de frecvenţă* $F'(x)=f(x)$.

Funcţia de frecvenţă este tot una cu densitatea de probabilitate din teoria probabilităţilor.

Avem

$$F(x)=\int_{-\infty}^{x} f(t)\,dt.$$

Distribuţiile statistice sînt definite fie prin legea de probabilitate respectivă, fie prin funcţia de distribuţie,

Capitolul II

Momente

§ 1. Definiţia momentelor şi valorilor medii

Să considerăm distribuţia discretă:

$$X\begin{pmatrix} x_k \\ s_k \end{pmatrix} \quad (k=1,\ 2,\ \ldots,\ n),$$

unde am notat prin x_k şi s_k, respectiv, valoarea argumentului şi frecvenţa relativă.

Prin definiţie, se numeşte *moment de ordinul p* al distribuţiei, expresia

$$M_p=\sum_{k=1}^{n} s_k x_k^p. \tag{II.1,1}$$

In cazul unei distribuţii continue, expresia dată de (II.1,1) devine

$$M_p=\int_{-\infty}^{+\infty} x^p f(x)\,dx. \tag{II.1,2}$$

Dacă variabila ia valori cuprinse într-un interval finit (α, β), înseamnă că pentru $x<\alpha$ şi $x>\beta$, putem lua $f(x)=0$, deci putem lua şi atunci ca limite ale integralei $-\infty$ şi $+\infty$.

Scanned with OKEN Scanner

Folosind integrala Stieltjes, momentul de ordinul p al unei distribuții oarecare este dat de expresia

$$M_p = \int_{-\infty}^{+\infty} x^p dF(x), \qquad \text{(II.1,3)}$$

unde $F(x)$ este funcția de distribuție, dacă această integrală există. Dacă $F(x)$ este funcția de distribuție a unei distribuții cu o funcție de frecvență $f(x)$ continuă, atunci și integrala Stieltjes devine integrală ordinară. In acest caz, expresia (II.1,2) a momentului M_p se prezintă ca un caz particular al expresiei (II.1,3).

Acum este de la sine înțeles că, dacă $F(x)$ este funcția de distribuție a unei distribuții discrete, expresia (II.1,3) a momentului M_p cuprinde de asemenea ca un caz particular și expresia (II. 1,1), deoarece integrala Stieltjes se reduce la o sumă obișnuită pentru valori discontinue ale variabilei.

Momentul absolut de ordin p este, prin definiție, în cazul unui număr finit de valori $\overline{M}_p = \sum_{k=1}^{m} s_k |x_k|^p$, iar în cazul unei distribuții oarecare $F(x)$: $\overline{M}_p = \int_{-\infty}^{+\infty} |x|^p dF(x)$ dacă există.

Momentele au o importanță deosebită, deoarece; în anumite condiții, ele caracterizează distribuția. Fiind dată succesiunea de numere

$$M_0 = 1,\ M_1,\ M_2, \ldots, M_n, \ldots,$$

problema momentelor constă în determinarea condițiilor în care aceste numere sînt momentele unei distribuții și construirea ei. Problema a fost pusă de Stieltjes (1894) și reluată de H. Hamburger (1920) și G. Castelnuovo (1930), care au reușit să-i dea o soluție definitivă.

Valori medii. Momentul de ordin 1 al variabilei x cu distribuția $F(x)$ sau, cum spunem foarte des, al distribuției $F(x)$, este identic cu valoarea medie a acestei variabile

$$M_1 = m_1 = M(x) = \int_{-\infty}^{+\infty} x\, dF(x),$$

în cazul general, sau

$$M_1 = m_1 = \sum_{k=1}^{m} S_k x_k$$

pentru o distribuţie cu un număr finit de valori. De aceea momentul de ordin p al distribuţiei $F(x)$ poate fi considerat ca media variabilei x^p:

$$M_p = M(x^p).$$

Media de ordinul p, pentru p întreg *mai mare* ca unu este

$$m_p = (\overline{M_p})^{\frac{1}{p}}.$$

Este evident că pentru valorile pare ale lui p avem $\overline{M_p} = M_p$ prin urmare

$$m_{2n} = (M_{2n})^{\frac{1}{2n}}. \qquad (n = 1, 2, \ldots).$$

Definiţia de mai sus se poate generaliza în diferite feluri. Aşa, de pildă, în cazul unei distribuţii discrete pentru care valorile x_k sînt toate pozitive

$$m_{-1} = (M_{-1})^{-1}$$

ne dă *media armonică* pentru care putem scrie

$$\frac{1}{m_{-1}} = \sum_{i=1}^{n} \frac{s_i}{x_i}.$$

De asemenea în cazul cînd $x_k > 0$ $(k = 1, 2, \ldots, n)$ putem atribui mediei de ordinul zero, prin definiţie, valoarea

$$\lim (M_p)^{\frac{1}{p}} = m_0.$$

Valoarea ce se obţine ca limită este *media geometrică*

$$m_0 = x_1^{s_1} x_2^{s_2} \ldots x_n^{s_n}.$$

In sfîrşit, pentru $p = 2$, avem *valoarea medie pătratică*, foarte des întrebuinţată în aplicaţii:

$$\sqrt{\sum_{k=1}^{n} s_k x_k^2} = \sqrt{s_1 x_1^2 + s_2 x_2^2 + \ldots + s_n x_n^2}.$$

Scanned with OKEN Scanner

Să observăm că primele două momente sînt legate de inegalitatea

$$\sqrt{M_2} \geq |M_1|,$$

care se demonstrează folosind inegalitatea lui Lagrange

$$\sum a_k^2 \sum b_k^2 - (\sum a_k b_k)^2 = \sum (a_k b_l - a_l b_k)^2,$$

unde punem $a_k = \sqrt{s_k}$, $b_k = \sqrt{s_k} \cdot x_k$ și ținem seama că $\sum_{k=1}^{n} s_k = 1$.

Avem, așa dar, pentru valori medii relația

$$m_2 > |m_1|.$$

In practică, pentru caracterizarea unei distribuții este suficientă, de cele mai multe ori, cunoașterea primelor patru sau cinci momente.

Exemplul 1. Distribuția binomială corespunzătoare la n probe succesive independente în care evenimentul A are probabilitatea constantă p dă loc următoarei distribuții

$$X: \begin{pmatrix} i \\ C_n^i p^i q^{n-i} \end{pmatrix} \qquad (i = 0,\ 1,\ 2, \ldots,\ n).$$

Pentru calcularea momentelor acestei distribuții vom folosi expresia (II.1,1)

$$M_j = \sum_{i=0}^{n} i^j C_n^i p^i q^{n-i} \qquad (j = 1,\ 2,\ 3, \ldots).$$

Calcularea sumelor din partea a doua se face ușor, derivînd succesiv în raport cu p expresia $\sum C_n^i p^i q^{n-i}$. Obținem astfel

$$M_1 = p \frac{\partial}{\partial p} (p+q)^n = np\,(p+q)^{n-1} = np,$$

$$M_2 = p \frac{\partial}{\partial p} (p+q)^n + p^2 \frac{\partial^2}{\partial p^2} (p+q)^n = np\,(p+q)^{n-1} +$$

$$+ n(n-1)\,p^2\,(p+q)^{n-2} = np + n(n-1)\,p^2 = np^2 + npq,$$

$$M_3 = p \frac{\partial}{\partial p} (p+q)^n + 3p^2 \frac{\partial^2}{\partial p^2} (p+q)^n + p^3 \frac{\partial^3}{\partial p^3} (p+q)^n =$$

$$= np + 3n(n-1)\,p^2 + n(n-1)(n-2)\,p^3 = n^3p^3 + 3n^2p^2q + npq\,(1-2p),$$

. . . . . .

Momentele acestei distribuţii se pot obţine şi pe altă cale, folosind proprietăţile sumei şi produsului de variabile. In proba de rang l avem distribuţia următoare

$$X_l \begin{pmatrix} 1 & 0 \\ p & q \end{pmatrix}.$$

Suma valorilor variabilei obţinute în cele n probe, corespunzătoare distribuţiei X se obţine însumînd valorile corespunzătoare fiecăreia dintre distribuţiile X_l. Atunci, putem scrie

$$M_1(X)=M\left(\sum_{l=1}^{n} X_l\right)=\sum_{l=1}^{n} M(X_l)=np,$$

$$M_2(X)=M\left(\sum_{l=1}^{n} X_l\right)^2=\sum_{l=1}^{n} M(X_l^2)+2\sum_{i,j=1}^{n} M(X_i)\,M(X_j)=$$

$$=np+n(n-1)p^2=n^2p^2+npq,$$

$$M_3(X)=M\left(\sum_{l=1}^{n} X_l\right)^3=\sum_{l=1}^{n} M(X_l^3)+3\sum_{i,j=1}^{n} M(X_l^2)\,M(X_j)+$$

$$+6\sum_{i,j,k=1}^{n} M(X_l)\,M(X_j)\,M(X_k)=np+3n(n-1)p^2+n(n-1)(n-2)p^3=$$

$$=n^3p^3+3n^2p^2q+npq(1-2p),$$

. . . . . .

Vom expune pe larg modul de calcul al momentelor distribuţiei binomiale în cap. V, § 1.

Exemplul 2. Distribuţia normală are funcţia de frecvenţă

$$f(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}.$$

Momentele ei pare sînt date de formula (II.1,2)

$$M_{2k}=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} x^{2k}e^{-\frac{x^2}{2}}dx,$$

iar cele impare sînt nule

$$M_{2k+1}=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} x^{2k+1}e^{-\frac{x^2}{2}}dx=0,$$

deoarece funcţia $x^{2k+1}e^{-\frac{x^2}{2}}$ este impară. Rămîne să determinăm numai momentele de ordin par. Integrînd prin părţi, avem

$$M_{2k}=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}x^{2k-1}d\left(-e^{-\frac{x^2}{2}}\right)=-\frac{1}{\sqrt{2\pi}}\left[\left(x^{2k-1}e^{-\frac{x^2}{2}}\right)_{-\infty}^{+\infty}-\right.$$

$$\left.-(2k-1)\int_{-\infty}^{+\infty}x^{2k-2}e^{-\frac{x^2}{2}}dx\right]=(2k-1)\cdot\int_{-\infty}^{+\infty}x^{2k-2}e^{-\frac{x^2}{2}}dx.$$

Deducem relaţiile de recurenţă

$$M_{2k}=(2k-1)\,M_{2k-2},$$

$$M_{2k-2}=(2k-3)\,M_{2k-4},$$

$$M_{2k-4}=(2k-5)\,M_{2k-6},$$

$$\cdots\cdots\cdots$$

$$M_4=3M_2.$$

$$M_2=M_0.$$

Inmulţind membrii din stînga şi din dreapta ai acestor egalităţi şi înlocuind pe M_0 cu 1, obţinem

$$M_{2k}=1\cdot3\cdot5\ldots(2k-3)\,(2k-1)$$

sau

$$M_{2k}=\frac{(2k)!}{2^k\cdot k!}.$$

§ 2. Inegalităţi între valori medii şi momente

Valorile medii şi momentele unei distribuţii statistice satisfac cîteva inegalităţi importante, care rezultă din teoremele următoare, *demonstrate în presupunerea că diversele momente considerate există.*

Teoremă. *Oricare ar fi numărul întreg m, valoarea medie* $\sqrt[2m]{M_{2m}}$ *satisface inegalităţile*

$$A\leq\sqrt[2m]{M_{2m}}\leq B,$$

unde am notat cu A şi B respectiv cea mai mică şi cea mai mare dintre valorile absolute ale variabilei x.

Intr-adevăr, prin definiţie avem

$$\sqrt[2m]{M_{2m}}=\left[\int_{-\infty}^{+\infty} x^{2m}\,dF(x)\right]^{\frac{1}{2m}}. \qquad \text{(II.2,1)}$$

Cum valorile absolute ale lui x sînt cuprinse între A şi B, putem scrie

$$x^{2m}\geqq A^{2m},\quad x^{2m}\leqq B. \qquad \text{(II.2,2)}$$

Inlocuind în (II.2,1) pe x^{2m} cu valorile rezultate din (II.2,2), găsim

$$\sqrt[2m]{M_{2m}}\geqq\left[A^{2m}\int_{-\infty}^{+\infty} dF(x)\right]^{\frac{1}{2m}}=A\left[\int_{-\infty}^{+\infty} dF(x)\right]^{\frac{1}{2m}}=A,$$

$$\sqrt[2m]{M_{2m}}\leqq\left[B^{2m}\int_{-\infty}^{+\infty} dF(x)\right]^{\frac{1}{2m}}=B\left[\int_{-\infty}^{+\infty} dF(x)\right]^{\frac{1}{2m}}=B.$$

deci

$$A\leqq\sqrt[2m]{M_{2m}}\leqq B.$$

Teoremă. *Valoarea medie* m_m *este crescătoare*

$$m_m\leqq m_{m+1}$$

oricare ar fi numărul întreg $m\geqq 1$ *pentru* $x>0$.
Să considerăm expresia

$$\int_{-\infty}^{+\infty}\left(u|x|^{\frac{m-1}{2}}+v|x|^{\frac{m+1}{2}}\right)^2 dF(x)\geqq 0,$$

oricare ar fi numerele u şi v.
Dezvoltînd pătratul de sub semnul integrală şi ţinînd seama de definiţia momentelor, avem

$$u^2M_{m-1}+2uv\,M_m+v^2M_{m+1}\geqq 0.$$

Această expresie fiind o formă definită pozitivă, discriminantul ei trebuie să fie pozitiv

$$\begin{vmatrix} M_{m-1} & M_m \\ M_m & M_{m+1}\end{vmatrix}=M_{m-1}\cdot M_{m+1}-M_m^2\geqq 0,$$

sau

$$M_m^2 \leq M_{m-1} \cdot M_{m+1}.$$

Observăm că pentru $m=1$, regăsim proprietatea valorii medii pătratice, despre care am vorbit în paragraful precedent.

Din ridicarea la puterea m în ultima inegalitate, dînd valori întregi succesive lui m, obţinem următoarele relaţii de recurenţă

$$M_m^{2m} \leq M_{m-1}^m \cdot M_{m+1}^m,$$

$$M_{m-1}^{2m-2} \leq M_{m-2}^{m-1} \cdot M_m^{m-1},$$

$$M_{m-2}^{2m-4} \leq M_{m-3}^{m-2} \cdot M_{m-1}^{m-2},$$

$$\cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot$$

$$M_4^8 \leq M_3^4 \cdot M_5^4,$$

$$M_3^6 \leq M_2^3 \cdot M_4^3,$$

$$M_2^4 \leq M_1^2 \cdot M_3^2,$$

$$M_1^2 \leq M_0 \cdot M_2.$$

Inmulţind membru cu membru şi făcînd toate reducerile, rezultă

$$M_m^{m+1} \leq M_{m+1}^m.$$

Ridicînd acum această inegalitate la puterea $\frac{1}{m(m+1)}$, deducem inegalitatea

$$\sqrt[m]{M_m} \leq \sqrt[m+1]{M_{m+1}},$$

care reprezintă tocmai teorema enunţată.

Teorema se poate generaliza, pentru orice $m>1$. Demonstrarea ei se poate face pe baza *inegalităţii lui Hölder-Iensen*

$$\varphi\left[\frac{\int_{-\infty}^{+\infty} |x|\, dF(x)}{\int_{-\infty}^{+\infty} dF(x)}\right] \leq \frac{\int_{-\infty}^{+\infty} \varphi(x)\, dF(x)}{\int_{-\infty}^{+\infty} dF(x)},$$

unde $\varphi(x)$ este o funcţie convexă regulată de x, pentru $x>0$ şi integrala este luată în sensul integralei Stieltjes.

In particular, să luăm $\varphi(x)=x^p$, care este o funcție convexă regulată pentru $x>0$ și $p>1$. In acest caz, inegalitatea lui Hölder-Iensen, devine

$$\left[\frac{\int_{-\infty}^{+\infty}|x|\,dF(x)}{\int_{-\infty}^{+\infty}dF(x)}\right]^p \leqslant \frac{\int_{-\infty}^{+\infty}|x|^p\,dF(x)}{\int_{-\infty}^{+\infty}dF(x)}$$

sau

$$\left[\int_{-\infty}^{+\infty}|x|\,dF(x)\right]^p \leqslant \int_{-\infty}^{+\infty}|x|^p\,dF(x),$$

deoarece

$$\int_{-\infty}^{+\infty}dF(x)=1.$$

Se înțelege că această inegalitate rămîne valabilă și pentru valorile $|x|^m$, întrucît ea are loc oricare ar fi $x>0$

$$\left[\int_{-\infty}^{+\infty}|x|^m dF(x)\right]^p \leqslant \int_{-\infty}^{+\infty}|x|^{mp}\,dF(x). \qquad \text{(II.2,3)}$$

Ridicînd la puterea $\frac{1}{mp}$, obținem

$$\sqrt[m]{\overline{M}_m} \leqslant \sqrt[mp]{\overline{M}_{mp}},$$

oricare ar fi $m>0$ și $p>1$.
Ultima inegalitate ne arată că

$$\sqrt[m]{\overline{M}_m} \leqslant \sqrt[m']{\overline{M}_{m'}} \qquad (0<m<m').$$

Să examinăm și cazul $m<0$, în ipoteza că inegalitatea lui Hölder-Iensen are loc, evident, și pentru valorile $|x|^{-r}$ $(r>0)$, cu presupunerea implicită că $\int_{-a}^{+b}|x|^{-r}\,dF(x)$ există pentru orice $a>0$, $b>0$.
Inlocuind atunci $|x|^m$ prin $|x|^{-r}$ în (II.2,3), avem

$$\left[\int_{-\infty}^{+\infty}|x|^{-r}dF(x)\right]^p \leqslant \int_{-\infty}^{+\infty}|x|^{-pr}\,dF(x).$$

Ridicînd și aici la puterea $-\frac{1}{pr}$, deducem

$$\sqrt[-r]{\bar{M}_{-r}} \geq \sqrt[-pr]{\bar{M}_{-pr}}.$$

Punem $m=-pr$ și $m'=-r$; rezultă

$$\sqrt[m]{\bar{M}_m} \leq \sqrt[m']{\bar{M}_{m'}} \qquad (m<m'<0).$$

Teorema este astfel demonstrată pentru orice valori reale ale lui m.

In continuare, vom menționa unele inegalități mai însemnate între momente.

Teoremă. *Oricare ar fi r, momentul M_r verifică întotdeauna inegalitățile*

$$A \leq M_r \leq B,$$

A și B fiind marginea inferioară și cea superioară ale valorilor lui x.

Demonstrația este simplă și imediată.

Putem scrie

$$M_r = \int_{-\infty}^{+\infty} x^r \, dF(x) \geq A \int_{-\infty}^{+\infty} dF(x) = A.$$

In mod similar avem

$$M_r \leq B.$$

Inegalitățile se transformă în egalități cînd distribuția r este concentrată într-un singur punct.

Teoremă. *Pentru orice întreg k $(0, 1, 2, \ldots, n, n+1, \ldots, 2n)$, momentele M_k, presupuse finite, verifică inegalitatea*

$$\begin{vmatrix} M_0 & M_1 & M_2 \ldots M_n \\ M_1 & M_2 & M_3 \ldots M_{n+1} \\ M_2 & M_3 & M_4 \ldots M_{n+2} \\ \cdots & \cdots & \cdots \\ M_n & M_{n+1} & M_{n+2} \ldots M_{2n} \end{vmatrix} \geq 0.$$

Fie forma pătratică definită pozitivă

$$\Phi = \int_{-\infty}^{+\infty} (\lambda_0 + \lambda_1 x + \lambda_2 x^2 + \ldots + \lambda_n x^n)^2 \, dF(x) \geqq 0.$$

Adică

$$\Phi = \sum_{i,j}^{0,1,\ldots,n} \lambda_i \lambda_j \int_{-\infty}^{+\infty} x^{i+j} \, dF(x) \geqq 0.$$

Ţinînd seamă de definiţia momentului avem

$$\Phi = \sum_{i,j=0}^{n} M_{i+j} \lambda_i \lambda_j \geqq 0. \qquad \text{(II.2,4)}$$

Φ fiind o formă pătratică definită pozitivă, discriminantul său este pozitiv şi astfel teorema este demonstrată.
Considerînd matricea de ordinul N

$$\left\| \begin{array}{llll} M_0 & M_1 & M_2 \ldots & M_N \\ M_1 & M_2 & M_3 \ldots & M_{N+1} \\ M_2 & M_3 & M_4 \ldots & M_{N+2} \\ \cdots & \cdots & \cdots & \cdots \\ M_N & M_{N+1} & M_{N+2} \ldots & M_{2N} \end{array} \right\|,$$

rezultă din această teoremă că determinanţii următori sînt toţi pozitivi

$$M_0 \geqq 0, \quad \begin{vmatrix} M_0 & M_1 \\ M_1 & M_2 \end{vmatrix} \geqq 0, \ldots, \quad \begin{vmatrix} M_0 & M_1 \ldots M_N \\ M_1 & M_2 \ldots M_N \\ \cdots & \cdots \\ M_N & M_{N+1} \ldots M_{2N} \end{vmatrix} \geqq 0,$$

pentru orice N.

§ 3. Momente centrate

In § 1 al acestui capitol am definit ca moment de ordinul p al unei distribuţii expresia

$$M_p = \int_{-\infty}^{+\infty} x^p \, dF(x),$$

dacă integrala din membrul al doilea există.

Dacă în expresia momentului M_p, raportăm valorile variabilei x la o origine luată în punctul de abscisă M_1 (momentul de ordinul întîi) obţinem *momentul centrat de ordinul p* al distribuţiei, a cărui expresie este

$$m_p = \int_{-\infty}^{+\infty} (x - M_1)^p \, dF(x), \qquad \text{(II.3.1)}$$

care pentru o distribuţie finită x se reduce la

$$m_p = \sum_{k=1}^{n} s_k (x_k - M_1)^p.$$

In cele ce urmează vom nota momentul de ordinul întîi, deseori numai prin M.

Dezvoltînd în (II.3,1), avem expresia momentului centrat de ordinul p, m_p, în funcţie de momentele obişnuite M_r $(r = 1, 2, \ldots, p)$ ale aceleiaşi distribuţii

$$m_p = M_p - C_p^1 M_{p-1} M_1 + C_p^2 M_{p-2} \cdot M_1^2 - \ldots + (-1)^{p-1} C_p^p M_1^p. \qquad \text{(II.3,2)}$$

Particularizînd valorile lui p în relaţia (II.3,2), rezultă următoarele valori pentru momentele centrate de diferite ordine

$$m_1 = 0,$$

$$m_2 = M_2 - M_1^2,$$

$$m_3 = M_3 - 3M_2 M_1 + 2M_1^3,$$

$$m_4 = M_4 - 4M_3 M_1 + 6M_2 M_1^2 - 3M_1^4,$$

. . . . . .

Folosind aceste relaţii putem acum găsi uşor momentele centrate ale distribuţiei binomiale şi normale, ale căror momente obişnuite au fost calculate în § 1 al capitolului de faţă.

Obţinem astfel, pentru momentele centrate ale distribuţiei binomiale,

$$m_1 = 0,$$
$$m_2 = n^2 p^2 + npq - n^2 p^2 = npq,$$
$$m_3 = npq\ (q - p),$$
$$m_4 = 3p^2 q^2 n^2 + pqn\ (1 - 6pq),$$
. . . . . .

Scanned with OKEN Scanner

Pentru distribuţia normală, deoarece $M_1=0$, avem

$$m_{2r}=M_{2r}=\frac{(2r)!}{2^r\cdot r!}, \quad m_{2r+1}=M_{2r+1}=0.$$

Momentele centrate se bucură de următoarea proprietate de invarianţă care le face utile în aplicaţii.

Teoremă. *Momentele centrate m_r ale oricărei distribuţii statistice sînt independente faţă de originea aleasă* ($r=1, 2, \ldots$).

Intr-adevăr, putem scrie

$$m_p=\int_{-\infty}^{+\infty}(x-M)^p\,dF(x),$$

$$M=\int_{-\infty}^{+\infty}x\,dF(x).$$

Făcînd schimbarea de origine în punctul de abscisă a, deducem

$$m_p(a)=\int_{-\infty}^{+\infty}[(x-a)-M(a)]^p\,dF(x),$$

unde, înlocuind expresia

$$M(a)=\int_{-\infty}^{+\infty}(x-a)\,dF(x)=M-a,$$

obţinem în cele din urmă

$$m_p(a)=\int_{-\infty}^{+\infty}[(x-a)-(M-a)]^p\,dF(x)=m_p.$$

Inegalitatea lui Cebîşev. Momentul funcţiei $g(x)$ în raport cu distribuţia $F(x)$ este, prin definiţie,

$$\int_{-\infty}^{+\infty}g(x)\,dF(x), \tag{II.3,3}$$

unde $g(x)$ este o funcţie de x, dacă, bineînţeles integrala există. Dacă în expresia (II.3,3) ne limităm numai la valorile $|g(x)|\geqslant K$, găsim

$$\int_{-\infty}^{+\infty}|g(x)|\,dF(x)\geqslant K\int_{A}dF(x), \tag{II.3,4}$$

A fiind mulţimea punctelor ale căror abscise x satisfac condiţia

$$|g(x)| \geq K.$$

Din definiţia funcţiei $F(x)$ rezultă că

$$\int_A dF(x) = P[|g(x)| \geq K],$$

iar din inegalitatea (II.3,4) deducem, atunci, inegalitatea lui Cebişev

$$P[|g(x)| \geq K] \leq \frac{\int_{-\infty}^{+\infty} |g(x)|\, dF(x)}{K} \qquad \text{(II.3,5)}$$

sau

$$P[|g(x)| < K] \geq 1 - \frac{\int_{-\infty}^{+\infty} |g(x)|\, dF(x)}{K}. \qquad \text{(II.3,6)}$$

Considerînd pe $g(x)$ ca variabilă, $g(x) - M[g(x)]$ se numeşte *abaterea* variabilei $g(x)$. Momentul de ordinul al doilea al abaterii variabilei $g(x)$, adică $M_2[g - M(g)]$ se mai numeşte *dispersie*, iar valoarea medie a aceleiaşi abateri, $+\sqrt{M_2[g - M(g)]}$, mai poartă numele de *abatere medie pătratică* a variabilei $g(x)$.

Dacă $g(x) = (x - M)^2$, atunci dispersia $m_2 = \int_{-\infty}^{+\infty} (x - M)^2\, dF(x)$ este cunoscută în terminologia statistică sub numele de *varianţă* şi se notează cu σ^2, iar numărul pozitiv σ se numeşte *deviaţie standard*. Luînd acum în formele (II.3,5) şi (II.3,6) ale inegalităţii lui Cebîşev, $g(x) = (x - M)^2$ şi numărul pozitiv $K = h^2\sigma^2$, obţinem următoarele forme, folosite de obicei în aplicaţii pentru această inegalitate

$$P[|x - M| \geq h\sigma] \leq \frac{\sigma^2}{h^2\sigma^2} = \frac{1}{h^2},$$

$$P[|x - M| < h\sigma] \geq 1 - \frac{1}{h^2}.$$

Exemplu. Pentru $h=5$,

$$P[|x-M|\geq 5\sigma]\leq \frac{1}{25},$$

$$P[|x-M|<5\sigma]\geq \frac{24}{25}.$$

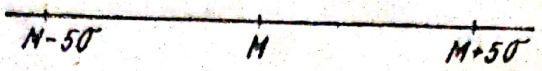


Fig. II.3,1

Cu alte cuvinte, cu o probabilitate mai mică decît $\frac{1}{25}$, valorile lui x se vor găsi în afara intervalului $(M-5\sigma,\ M+5\sigma)$ şi cuprinse în acest interval cu o probabilitate între $\frac{24}{25}$ şi 1 (fig. II.3,1).

§ 4. Funcţie caracteristică

Prin definiţie, *funcţia caracteristică este momentul de ordinul întîi al expresiei* e^{xti}, *pentru orice* t *real*

$$\varphi(t)=M_1(e^{xti})=M(e^{xti}). \tag{II.4,1}$$

Folosind integrala Stieltjes, funcţia caracteristică se scrie

$$\varphi(t)=\int_{-\infty}^{+\infty} e^{xti}\,dF(x), \tag{II.4,2}$$

care se reduce la o sumă obişnuită în cazul unei distribuţii discrete cu un număr finit sau infinit de valori ale variabilei

$$\varphi(t)=\sum_{k=1}^{n} s_k e^{x_k ti},\quad \varphi(t)=\sum_{k=1}^{\infty} s_k e^{x_k ti}$$

şi la o integrală ordinară în cazul unei distribuţii continue

$$\varphi(t)=\int_{-\infty}^{+\infty} e^{xti} f(x)\,dx.$$

Noţiunea de funcţie caracteristică prezintă avantajul că, în general, conduce la calcule mult mai simple faţă de cele obţinute cu ajutorul momentelor.

Funcţia caracteristică, determinată dacă este dată funcţia de distribuţie, la rîndul ei dacă este cunoscută, determină funcţia de distribuţie întocmai cum o determină succesiunea de momente *cînd ele există*.

Funcţia caracteristică se bucură de unele proprietăţi generale, care sînt o urmare imediată a definiţiei ei.

1) $$\varphi(0)=1.$$

Proprietatea rezultă imediat, dacă facem $t=0$ în expresiile (II.4,1) sau (II.4,2).

2) $|\varphi(t)| \leq \varphi(0)$.

3) $\varphi(t)$ este o funcţie continuă pentru orice t real, uniform continuă pe toată axa reală.

4) $\varphi(-t)=\overline{\varphi}(t)$,

unde am notat cu $\overline{\varphi}(t)$ conjugata funcţiei $\varphi(t)$.

5) Funcţia caracteristică este generatoare de momente în măsura în care ele există.

Cînd variabila aleatorie are un număr finit de valori, de pildă $a_1, a_2, \ldots, a_m$, cu probabilităţile respective $p_1, p_2, \ldots, p_m$, funcţia caracteristică este dată de seria

$$\varphi(t)=\sum_{j=1}^{m} p_j e^{ita_j}=\sum_{j=1}^{m} p_j \sum_{n=0}^{\infty} \frac{i^n t^n a_j^n}{n!}=$$

$$=\sum_{n=0}^{\infty} \frac{i^n}{n!}\left(\sum_j p_j a_j^n\right) t^n=\sum_{n=0}^{\infty} \frac{i^n}{n!} M_n t^n. \qquad \text{(II.4,3)}$$

Seria din membrul al doilea este absolut şi uniform convergentă ca o sumă de m serii absolut şi uniform convergente pentru orice t real.

Considerînd dezvoltarea în serie

$$\varphi(t)=\sum \frac{1}{n!}\left(\frac{d^n\varphi(t)}{dt^n}\right)_{t=0} t^n,$$

rezultă, prin comparaţie cu seria precedentă, că

$$i^n M_n=\left(\frac{d^n\varphi(t)}{dt^n}\right)_{t=0}. \qquad \text{(II.4,4)}$$

In cazul general al unei distribuţii $F(x)$ avem

$$\varphi(t)=\int_{-\infty}^{+\infty} e^{itx}dF(x),$$

cu condiţia să existe derivata de ordinul n

$$\frac{d^n\varphi(t)}{dt^n}=i^n\int_{-\infty}^{+\infty} e^{itx}x^n dF(x),$$

adică să existe momentul absolut de ordinul n

$$\overline{M}_n=\int_{-\infty}^{+\infty}|x|^n dF(x).$$

In acest caz există

$$M_n=\int_{-\infty}^{+\infty} x^n dF(x)$$

şi este legat de $\frac{d^n\varphi(t)}{dt^n}$ prin aceeaşi relaţie (II.4,4) ca şi în cazul distribuţiilor finite.

Reprezentarea funcţiei caracteristice prin seria (II.4,4) ai cărei coeficienţi sînt $\frac{i^nM_n}{n!}$ este valabilă dacă seria este convergentă şi anume pentru $|t|$ mai mic decît raza de convergenţă a seriei.

Fie $\varphi_1(t)$ funcţia caracteristică a variabilei X_1 şi $\varphi_2(t)$ funcţia caracteristică a variabilei X_2. Dacă variabilele X_1 şi X_2 sînt independente, funcţia caracteristică a sumei X_1+X_2 este $\varphi_1(t)\cdot\varphi_2(t)$. Proprietatea se păstrează pentru un număr finit oarecare de variabile $X_1, X_2, \ldots, X_s$ independente în ansamblul lor. Dacă funcţiile caracteristice respective sînt $\varphi_1(t), \varphi_2(t), \ldots, \varphi_s(t)$, funcţia caracteristică a variabilei $\sum_{j=1}^{s} X_j$ este

$$\varphi_1(t)\cdot\varphi_2(t)\ldots\varphi_s(t).$$

Exemplul 1. Funcţia caracteristică a distribuţiei binomiale. Fiecărei extracţii, oricare este rangul ei, îi corespunde distribuţia

$$\begin{pmatrix}1 & 0\\ p & q\end{pmatrix}.$$

Funcţia caracteristică este deci

$$\varphi(t)=pe^{ti}+q.$$

Dacă ţinem seama că funcţia caracteristică a unei sume de variabile independente este egală cu produsul funcţiilor caracteristice corespunzătoare, vom avea pentru un sistem de n extracţii succesive

$$\varphi(t)=\prod_{k=1}^{n}\varphi_k(t),$$

adică

$$\varphi(t)=(pe^{it}+q)^n.$$

De aici putem obţine fără greutate momentele M_k ($k=1, 2, \ldots$) ale distribuţiei binomiale, în baza relaţiei (II.4,4)

$$M_1=\frac{1}{i}\cdot\frac{d}{dt}[(pe^{it}+q)^n]_{t=0}=np,$$

$$M_2=\frac{1}{i^2}\frac{d^2}{dt^2}[(pe^{it}+q)^n]_{t=0}=np+n(n-1)p^2=n^2p^2+npq.$$

. . . . . .

Momentele centrate m_k ($k=1, 2, \ldots$) ale distribuţiei binomiale rezultă considerînd funcţia caracteristică a distribuţiei corespunzătoare abaterii variabilei. In acest caz, pentru fiecare extracţie de rang k, avem distribuţia

$$X_k'\begin{pmatrix}1-p & -p\\ p & q\end{pmatrix}.$$

Funcţia caracteristică a distribuţiei celor n extracţii succesive este evident

$$\varphi(t)=(pe^{qti}+qe^{-pti})^n.$$

Din această expresie a funcţiei caracteristice, deducem momentele centrate m_k ale distribuţiei binomiale, folosind aceleaşi relaţii ca mai sus

$$m_1=\frac{1}{i}\cdot\frac{d}{dt}[(pe^{qti}+qe^{-pti})^n]=n(pq-pq)=0,$$

$$m_2=\frac{1}{i^2}\cdot\frac{d^2}{dt^2}[(pe^{qti}+qe^{-pti})^n]=npq(p+q)=npq.$$

. . . . . .

Exemplul 2. Funcţia caracteristică a distribuţiei normale

$$\varphi(x)=\frac{1}{\sqrt{2\pi}}\cdot e^{-\frac{x^2}{2}}.$$

Prin definiţie, avem

$$\varphi(t)=M(e^{xti})=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}e^{xti}e^{-\frac{x^2}{2}}dx=e^{-\frac{t^2}{2}}.$$

Dezvoltînd în serie pe $e^{-\frac{t^2}{2}}$, obţinem următoarea expresie pentru funcţia caracteristică $\varphi(t)$:

$$\varphi(t)=1+\frac{t^2i^2}{2!}+1\cdot3\frac{t^4i^4}{4!}+1\cdot3\cdot5\frac{t^6i^6}{6!}+\ldots+1\cdot3\ldots(2k-1)\frac{t^{2k}i^{2k}}{(2k)!}+\ldots$$

Observăm că momentele de ordin impar, M_{2k+1} $(k=1, 2, \ldots)$ sînt nule, iar cele de ordin par sînt date de relaţia

$$M_{2k}=1\cdot3\cdot5\ldots(2k-3)(2k-1) \qquad (k=1, 2, \ldots).$$

Momentele centrate m_k $(k=1, 2, \ldots)$ coincid cu momentele obişnuite.

Funcţia generatoare. Inainte de a se întrebuinţa funcţia caracteristică s-a folosit, începînd de la Laplace, *funcţia generatoare*

$$g(t)=M(e^{xt})=\int_{-\infty}^{+\infty}e^{xt}dF(x),$$

cînd această funcţie există.

In cazul variabilelor cu un număr finit de valori, această funcţie există

$$g(t)=\sum_j p_j e^{a_j t}.$$

De cîte ori există, această funcţie este *reală.*

Avem şi în acest caz

$$g(0)=1,$$

precum şi dezvoltarea în serie de puteri cu coeficienţi proporţionali cu momentele

$$g(t)=\sum_{n=0}^{\infty}M_n\frac{t^n}{n!}.$$

Funcţia generatoare nu este mărginită, spre deosebire de funcţia caracteristică, dar adeseori este tot aşa de utilă ca şi aceasta.

Avem evident o funcţie generatoare şi pentru sisteme de r variabile.

Scanned with OKEN Scanner

§ 5. Cumulanți

Definiție. *Se numește cumulant de ordinul p numărul* $M_{[p]}$ *determinat de egalitatea*

$$i^p M_{[p]} = \left[\frac{d^{p-1}}{dt^{p-1}}\left[\frac{\varphi'(t)}{\varphi(t)}\right]\right]_{t=0} \qquad \text{(II.5,1)}$$

dacă există derivatele pînă la ordinul p ale funcției caracteristice $\varphi(t)$, deci dacă există momentele pînă la acest ordin, inclusiv. Egalitatea precedentă se poate scrie, încă, în aceleași condiții

$$i^p M_{[p]} = \frac{d^p}{dt^p}[\ln \varphi(t)],$$

cu următoarea observație:
Deoarece

$$\varphi(t) = R(t)\, e^{iw(t)+2ki\pi},$$

$\ln \varphi(t)$ este o funcție multiformă

$$\ln \varphi(t) = \ln R(t) + i[w(t) + 2k\pi].$$

Pe de altă parte, știm că $\varphi(0)=1$; există deci o determinare a argumentului $w(t)+2k\pi$ care pentru $t=0$ devine nulă. Această determinare, ce se notează cu $\omega(t)$, se numește argument principal în vecinătatea originii $t=0$ și vom înțelege că

$$\ln \varphi(t) = \ln R(t) + i\omega(t).$$

Formula (II.5,1) ne dă, rînd pe rînd, pentru $p=1, 2, 3, 4, \ldots$

$$iM_{[1]} = \frac{\varphi'(0)}{\varphi(0)} = \varphi'(0) = iM_1,$$
$$M_{[2]} = -\varphi''(0) + \varphi'^2(0) = M_2 - M_1^2,$$
$$M_{[3]} = M_3 - 3M_2M_1 + 2M_1^3,$$
$$M_{[4]} = M_4 - 4M_3M_1 - 3M_1^2 + 12M_2M_1^2 - 6M_1^4,$$

de unde și relațiile inverse

$$M_1 = M_{[1]},$$
$$M_2 = M_{[2]} + M_{[1]}^2,$$
$$M_3 = M_{[3]} + 3M_{[2]} \cdot M_{[1]} + M_{[1]}^3,$$
$$\ldots\ldots\ldots\ldots\ldots$$

Scanned with OKEN Scanner

Examinăm acum modul de transformare al cumulanţilor, şi în acelaşi timp al momentelor, prin transformarea

$$\xi = ax + b$$

efectuată asupra variabilei x. Această transformare revine la o schimbare a unităţii de măsură a lungimilor şi o schimbare a reperului. Dacă numim cu $\varphi(t)$ funcţia caracteristică a variabilei x şi cu $\psi(t)$ pe aceea a variabilei ξ, vom avea

$$\psi(t) = \int_{-\infty}^{+\infty} e^{i(ax+b)t}\, dF(x) = e^{ibt} \int_{-\infty}^{+\infty} e^{iat}\, dF(x) = e^{ibt}\, \varphi(at).$$

Rezultă

$$\begin{aligned}
\psi'(0) &= ib + a\varphi'(0),\\
\psi''(0) &= i^2b^2 + 2iab\varphi'(0) + a^2\varphi''(0),\\
\psi'''(0) &= i^3b^3 + 3i^2ab^2\varphi'(0) + 3ia^2b\varphi''(0) + a^3\varphi'''(0),\\
&\cdots\cdots\cdots\cdots\cdots
\end{aligned}$$

de unde *formulele de transformare ale momentelor*

$$\begin{aligned}
M_1' &= aM_1 - b,\\
M_2' &= a^2M_2 - 2abM_1 + b^2,\\
M_3' &= a^3M_3 - 3a^2bM_2 + 3ab^2M_1 - b^3,\\
&\cdots\cdots\cdots\cdots
\end{aligned}$$

precum şi *formulele de transformare ale cumulanţilor*

$$\begin{aligned}
M_{[1]}' &= aM_{[1]} - b,\\
M_{[2]}' &= a^2M_{[2]},\\
M_{[3]}' &= a^3M_{[3]},\\
&\cdots\cdots
\end{aligned}$$

ceea ce arată că

Începînd de la ordinul al doilea cumulanţii sînt invarianţi pentru schimbarea reperului.

Aceasta se vede şi direct din (II.5,1) deoarece

$$\frac{\psi'(t)}{\psi(t)} = ib + a\,\frac{\varphi'(at)}{\varphi(at)}.$$

Scanned with OKEN Scanner

Exemplul 1. *Cumulanții distribuției binomiale.*
Se observă că valorile găsite pentru cumulanții $M_{[k]}$ $(k=1, 2, 3)$ coincid cu momentele centrate m_k $(k=1, 2, 3)$ ale distribuției binomiale. Începînd de la cumulantul de ordinul patru, valorile respective nu mai coincid.

Exemplul 2. Pentru determinarea cumulanților $M_{[r]}(r=1,2, \ldots)$ a distribuției normale, plecăm de la funcția caracteristică corespunzătoare, ceea ce dă

$$\ln \varphi(t) = -\frac{t^2}{2}.$$

Aplicînd relația (II.5,1) obținem

$$M_{[1]} = \frac{1}{i} \cdot \frac{d}{dt} [\ln \varphi(t)]_{t=0} = [ti]_{t=0} = 0,$$

$$M_{[2]} = \frac{1}{i^2} \cdot \frac{d^2}{dt^2} [\ln \varphi(t)]_{t=0} = 1.$$

Este evident că toți cumulanții care urmează ($M_{[3]}$, $M_{[4]}, \ldots$) sînt nuli.

§ 6. Momente factoriale

Momentul factorial de ordinul p al distribuției $F(x)$ este definit de expresia

$$M_{(p)} = \int_{-\infty}^{+\infty} x(x-1)\ldots(x-p+1)\, dF(x). \qquad \text{(II.6,1)}$$

Introducînd notația simbolică

$$(x)_p = x(x-1)\ldots(x-p+1),$$

expresia (II.6,1) se mai scrie

$$M_{(p)} = \int_{-\infty}^{+\infty} (x)_p dF(x).$$

In cazul unei distribuții discrete cu n valori ale variabilei, integrala (II.6,1) se reduce la suma obișnuită

$$M_{(p)} = \sum_{k=1}^{n} (x_k)_p S_k.$$

In general, momentele factoriale se aplică la studiul distribuțiilor discrete sau la cele continue cu valori ale variabilei grupate pe

intervale. Din această cauză, deşi ele nu joacă un rol direct în statistica matematică, au însă o largă aplicaţie în domeniul proceselor de interpolare şi ajustare, auxiliare importante ale statisticii. De aceea vom examina aici unele proprietăţi ale momentelor factoriale.

Relaţiile între $(x)_p$ şi x^p se pot scrie în două moduri diferite:

In primul mod se exprimă $(x)_p$ ca un polinom în x^k $(k=0, 1, 2,..., p)$ cu ajutorul dezvoltării Mac Laurin

$$(x)_p=\sum_{k=1}^{p+1}\frac{1}{k!}\left(\frac{d^k(x)_p}{dx^k}\right)_{x=0}x^k. \qquad \text{(II.6,2)}$$

In al doilea mod se exprimă x^p prin polinoamele $(x)_k$ $(k=0, 1, 2,..., p)$ cu ajutorul formulei lui Newton.

$$x^p=\sum_{k=1}^{p+1}\frac{1}{k!}(\Delta^k x^p)_{x=0}(x)_k. \qquad \text{(II.6;3)}$$

Coeficienţii primei dezvoltări sînt *numerele* S_p^k *ale lui Stirling de prima speţă*

$$S_p^k=\frac{1}{k!}\left(\frac{d^k(x)_p}{dx^k}\right)_{x=0}. \qquad \text{(II.6,4)}$$

Coeficienţii dezvoltării (II.6,3) sînt *numerele lui Stirling de speţa a doua*

$$T_p^k=\frac{1}{k!}[\Delta^k x^p]_{k=0}. \qquad \text{(II.6,5)}$$

Este evident din (II.6,3) că

$$S_p^k=0, \quad T_p^k=0.$$

pentru $k>p$.

Din formulele precedente rezultă imediat următoarele două relaţii care exprimă momentul factorial $M_{(p)}$ prin momentele ordinare M_k $(k=1, 2,..., p)$ şi, reciproc, momentul ordinar M_p prin momentele factoriale $M_{(k)}$ $(k=1, 2,..., p)$. Aceste reprezentări au drept coeficienţi numerele lui Stirling S_p^k şi T_p^k:

$$M_{(p)}=\sum_{k=1}^{p}M_k S_p^k, \quad M_p=\sum_{k=1}^{p}M_{(k)}T_p^k.$$

Exemplul 1. *Momentele factoriale ale distribuției binomiale.*
Găsim, în baza primei teoreme, următoarele valori pentru momentele factoriale

$$M_{(1)} = S_1^1 M_1 = M_1 = np,$$

$$M_{(2)} = S_2^1 M_1 + S_2^2 M_2 = -M_1 + M_2 = -np + n^2p^2 + npq = n(n-1)p^2,$$

. . . . . .

De asemenea, în baza teoremei a doua, deducem

$$M_1 = T_1^1 M_{(1)} = M_{(1)} = np,$$

$$M_2 = T_2^1 M_{(1)} + T_2^2 M_{(2)} = M_{(1)} + M_{(2)} = np + n(n-1)p^2 = n^2p^2 + npq.$$

. . . . . .

Exemplul 2. *Momentele factoriale ale distribuției Poisson* care este o distribuție discontinuă de aproximare asimptotică a distribuției binomiale pentru valori mici ale lui p.
Funcția sa caracteristică este

$$\varphi(t) = M(e^{\lambda ti}) = \sum_{\lambda=0}^{\infty} \frac{a^\lambda}{\lambda!} e^{-a} \cdot e^{\lambda ti} = e^{-a} \sum_{\lambda=0}^{\infty} \frac{(ae^{ti})^\lambda}{\lambda!} = e^{a(e^{ti}-1)}.$$

De aici, rezultă momentele M_r $(r=1, 2, \ldots)$ ale distribuției Poisson

$$M_1 = \frac{1}{i} \frac{d}{dt} \left[e^{a(e^{ti}-1)}\right]_{t=0} = a,$$

$$M_2 = \frac{1}{i^2} \frac{d^2}{dt^2} \left[e^{a(e^{ti}-1)}\right]_{t=0} = a^2 + a,$$

. . . . . .

Obținem, astfel, valorile de mai jos pentru momentele factoriale ale aceleiași distribuții

$$M_{(1)} = S_1 M_1 = M_1 = a,$$

$$M_{(2)} = S_2^1 M_1 + S_2^2 M_2 = -M_1 + M_2 = a^2.$$

. . . . . .

De observat că momentele factoriale ale unei distribuții se pot obține considerînd expresia

$$\Phi(t) = \int_{-\infty}^{+\infty} t^x dF(x) \qquad \text{(II.6,6)}$$

care este funcția generatoare a momentelor factoriale.

Derivînd expresia (II.6,6) succesiv în raport cu t, avem

$$\frac{d}{dt}[\Phi(t)] = \int_{-\infty}^{+\infty} x t^{x-1} dF(x),$$

$$\frac{d^2}{dt^2}[\Phi(t)] = \int_{-\infty}^{+\infty} x(x-1)\, t^{x-2} dF(x), \qquad \text{(II.6,7)}$$

$$\cdots\cdots\cdots\cdots$$

$$\frac{d^{(r)}}{dt^r}[\Phi(t)] = \int_{-\infty}^{+\infty} x(x-1)\ldots(x-r+1)\, t^{x-r}\, dF(x).$$

Dacă, acum, facem în relaţiile (II.6,7) $t=1$, se vede uşor că obţinem momentele factoriale şi deci putem scrie

$$M_{(r)} = \frac{d^{(r)}}{dt^r}[\Phi(t)]_{t=1}. \qquad \text{(II.6,8)}$$

Cu ajutorul relaţiei (II.6,8) se determină direct momentele factoriale $M_{(r)}$ ale unei distribuţii, dacă funcţia $\Phi(t)$, împreună cu derivatele sale de orice ordin există. Această relaţie ne scuteşte de a mai calcula momentele factoriale ale unei distribuţii, folosind relaţia liniară în raport cu numerele lui Stirling.

§ 7. Corecţiile momentelor datorită grupării

Am arătat în cap. I, în legătură cu distribuţiile, că valorile variabilei unei distribuţii continue se grupează, adesea, după intervale egale. In aplicaţii pentru calculul momentelor de diferite ordine, toate frecvenţele corespunzătoare unui interval se raportează la mijlocul intervalului. Prin acest procedeu, înlocuim în realitate o variabilă continuă printr-o variabilă discontinuă, care poate lua numai valorile

$$x_k = x_0 + ka \qquad (k=0,\ \pm 1,\ \pm 2, \ldots),$$

unde am notat cu a lungimea intervalului, iar frecvenţa corespunzătoare punctului x_k este egală cu frecvenţa totală corespunzătoare intervalului $\left(x_k - \frac{a}{2},\ x_k + \frac{a}{2}\right)$.

Atunci, valorile x_k ale variabilei pot fi considerate drept centre ale intervalelor $\left(x_k-\frac{a}{2},\ x_k+\frac{a}{2}\right)$.

Lucrînd în felul acesta, aducem evident simplificări în calcule, dar în același timp se introduc erori sistematice prin concentrarea întregii mase de frecvențe în mijlocul intervalului.

Pentru înlăturarea acestui inconvenient, datorit erorilor care afectează în mod sistematic calculele diferitelor caracteristici ale distribuției, se fac unele corecții, numite *corecții Sheppard* (1898), formulate într-un mod mai precis de H. Wold (1934).

Insemnînd prin M_k, $M_{(k)}$ și $M_{[k]}$, respectiv momentul obișnuit, momentul factorial și cumulantul de ordinul k și prin M'_k, $M'_{(k)}$ și $M'_{[k]}$ respectiv aceleași caracteristici, calculate în baza grupării pe intervale, obținem următoarele relații care determină corecțiile Sheppard pentru distribuțiile a căror funcție de frecvență $f(x)$ se bucură de proprietatea că admite derivate continue pînă la ordinul $2s+1$ inclusiv și că există un număr întreg k, astfel încît $x^{k+2}f^{(i)}(x)$ este mărginită pentru orice valoare a lui x, oricare ar fi $i=0,\ 1,\ 2,\ \ldots,\ 2s$.

Aceste condiții sînt evident îndeplinite de clasa distribuțiilor pentru care avem $f(x)=0$ pentru $x<a$ și $f(x)=0$ pentru $x>b$, adică pentru distribuțiile a căror variabilă este cuprinsă într-un interval finit; de asemenea de distribuțiile normale.

Formulele generale stabilite de H. Wold sînt următoarele:

$$M_m=\sum_{i=0}^{m}C_m^iB_{m-i}\left(\frac{a}{2}\right)^{m-i}M_m,$$

$$M_{(m)}=\sum_{j=0}^{m}C_m^ja^j\left(B_j\left(\frac{3}{2}\right)\right)_{j+2}M'_{(m-j)},$$

$$M_{[m]}=M'_{[m]}-\frac{B_ma^m}{m},$$

unde B_k reprezintă numărul lui Bernoulli de ordinul k.

Aceste formule sînt valabile pentru toate momentele al căror ordin verifică relația $m\leq k$ și cu un grad de aproximație dat de a^{2s}. Se vede că ele sînt cu atît mai exacte cu cît a este mai mic.

Pentru distribuţiile îndeplinind condiţiile menţionate, avem următoarele valori ale primelor momente teoretice

$$M_1 = M'_1,$$
$$M_2 = M'_2 - \frac{a^2}{12},$$
$$M_3 = M'_3 - \frac{a^2}{4} M'_1,$$
$$M_4 = M'_4 - \frac{a^2}{2} M'_2 + \frac{7a^4}{240}, \qquad \text{(II.7,1)}$$
$$M_5 = M'_5 - \frac{5}{6} a^2 M'_3 + \frac{7}{48} a^4 M'_1,$$
$$M_6 = M'_6 - \frac{5}{4} a^2 M'_4 + \frac{7}{16} a^4 M'_2 - \frac{31}{1\,344} a^6,$$
. . . . . .

$$M_{(1)} = M'_{(1)},$$
$$M_{(2)} = M'_{(2)} - \frac{a^2}{12},$$
$$M_{(3)} = M'_{(3)} - \frac{a^2}{4} M'_{(1)} + \frac{a^3}{4}, \qquad \text{(II.7,2)}$$
$$M_{(4)} = M'_{(4)} - \frac{a^2}{2} M'_{(2)} + a^3 M'_{(1)} - \frac{71}{80} a^4,$$
$$M_{(5)} = M'_{(5)} - \frac{5}{6} a^2 M'_{(3)} + \frac{5}{2} a^3 M'_{(2)} - \frac{71}{16} a^4 M'_{(1)} + \frac{31}{8} a^5,$$
$$M_{(6)} = M'_{(6)} - \frac{5}{4} a^2 M'_{(4)} + 5a^3 M'_{(3)} - \frac{213}{16} a^4 M'_{(2)} + \frac{93}{4} a^5 M'_{(1)} - \frac{9\,129}{448} a^6,$$
. . . . . .

$$M_{[1]} = M'_{[1]},$$
$$M_{[2]} = M'_{[2]} - \frac{a^2}{12},$$
$$M_{[3]} = M'_{[3]}, \qquad \text{(II.7,3)}$$
$$M_{[4]} = M'_{[4]} + \frac{a^4}{120},$$
$$M_{[5]} = M'_{[5]},$$
$$M_{[6]} = M'_{[6]} - \frac{a^6}{252},$$
. . . . . .

La determinarea corecţiilor Sheppard, vom folosi unele teoreme din teoria calculului cu diferenţe finite
Să considerăm formula lui Euler-Mac-Laurin

$$\frac{1}{a}\int_{-\infty}^{+\infty} g(x)\,dx \int_{-\frac{a}{2}}^{\frac{a}{2}} f(x+u)\,du = \sum_{l=-\infty}^{+\infty} g(x_l) \int_{-\frac{a}{2}}^{+\frac{a}{2}} f(x_l+u)\,du + R \quad \text{(II.7,4)}$$

Restul R din formula Euler-Mac-Laurin are o valoare bine cunoscută. Menţionăm că, dacă în formula de mai sus $g(x)$ este un polinom de gradul maxim k, atunci, în virtutea faptului că $x^k f^{(i)}(x)$ $(i=1, 2, \ldots, 2s)$ este mărginită, rezultă după calcule simple

$$g(x) \int_{-\frac{a}{2}}^{\frac{a}{2}} f(x+u)\,du = O(x^{-2}),$$

pentru $x \to \pm\infty$.
Cum restul R este de forma

$$R=(-1)^{s+1}a^{2s+1}\int_{-\infty}^{+\infty} G(x)\left[\, g(x_0+ax)\int_{-\frac{a}{2}}^{\frac{a}{2}} f(x_0+ax+u)\,du \,\right]^{2s+1} dx,$$

unde $G(x)$ este o funcţie mărginită, oricare ar fi x, reiese, ţinînd seama de valoarea asimptotică a lui x, că $|R| = O(a^{2s})$.
Dacă $g(x)$ este un polinom de gradul maxim k, caz care se va întîlni pentru rezolvarea problemei noastre, gradul de aproximaţie în formula lui Euler-Mac-Laurin este de ordinul a^{2s}. Vom reţine acest lucru, neglijînd restul R în calculele care urmează.
Să luăm

$$\varphi(x) = \frac{1}{a}\int_{-\frac{a}{2}}^{\frac{a}{2}} g(x+u)\,du. \quad \text{(II.7,5)}$$

Derivînd, obţinem

$$\varphi'(x) = \frac{1}{a}\int_{-\frac{a}{2}}^{\frac{a}{2}} g'(x+u)\,du = \frac{1}{a}\left[g\left(x+\frac{a}{2}\right) - g\left(x-\frac{a}{2}\right)\right].$$

Punem în locul lui x, $x+\frac{a}{2}$ şi egalitatea devine

$$\Delta_a g(x)=\varphi'\left(x+\frac{a}{2}\right), \qquad \text{(II.7,6)}$$

unde

$$\Delta_a g(x)=\frac{g(x+a)-g(x)}{a}.$$

Integrînd prin părţi în membrul întîi al formulei Euler-Mac-Laurin (II.7,4), deducem

$$\frac{1}{a}\int_{-\infty}^{+\infty} g(x)\,dx \int_{-\frac{a}{2}}^{\frac{a}{2}} f(x+u)\,du=\left[\int_0^x g(t)\,dt\,\frac{1}{a}\int_{-\frac{a}{2}}^{\frac{a}{2}} f(x+u)\,du\right]_{-\infty}^{+\infty}-$$

$$-\frac{1}{a}\int_{-\infty}^{+\infty}\int_0^x g(t)\,dt\left[f\left(x+\frac{a}{2}\right)-f\left(x-\frac{a}{2}\right)\right]dx.$$

Dacă avem îndeplinită şi condiţia

$$\lim_{x\to\pm\infty}\left[\int_0^x g(t)\,dt\right]\left[\int_{-\frac{a}{2}}^{\frac{a}{2}} f(x+u)\,du\right]=0, \qquad \text{(I)}$$

atunci

$$\frac{1}{a}\int_{-\infty}^{+\infty} g(x)\,dx \int_{-\frac{a}{2}}^{\frac{a}{2}} f(x+u)\,du=$$

$$=-\frac{1}{a}\int_{-\infty}^{+\infty}\int_0^x g(t)\,dt\left[f\left(x+\frac{a}{2}\right)-f\left(x-\frac{a}{2}\right)\right]dx=$$

$$=-\frac{1}{a}\int_{-\infty}^{+\infty}\int_0^x g(t)\,f\left(x+\frac{a}{2}\right)dt\,dx+\frac{1}{a}\int_{-\infty}^{+\infty}\int_0^x g(t)\,f\left(x-\frac{a}{2}\right)dt\,dx.$$

In prima parte a ultimei expresii facem schimbarea de variabilă $y=x+\frac{a}{2}$, iar în partea a doua, schimbarea de variabilă $y=x-\frac{a}{2}$.

Scanned with OKEN Scanner

Rezultă

$$\frac{1}{a}\int_{-\infty}^{+\infty} g(x)\,dx \int_{-\frac{a}{2}}^{\frac{a}{2}} f(x+u)\,du = -\frac{1}{a}\int_{-\infty}^{+\infty}\int_{y-\frac{a}{2}}^{0} g(t)\,f(y)\,dt\,dy +$$

$$+\frac{1}{a}\int_{-\infty}^{+\infty}\int_{0}^{y+\frac{a}{2}} g(t)\,f(y)\,dt\,dy = \frac{1}{a}\int_{-\infty}^{+\infty} f(y)\,dy \int_{y-\frac{a}{2}}^{y+\frac{a}{2}} g(t)\,dt.$$

Ținînd seama acum de relațiile (II.7,4) și (II.7,5) avem egalitatea

$$\int_{-\infty}^{+\infty} \varphi(x)\,f(x)\,dx = \sum g(x_i) \int_{-\frac{a}{2}}^{\frac{a}{2}} f(x_i+u)\,du. \qquad \text{(II.7,7)}$$

Să ne alegem

1) $$\varphi(x) = x^m.$$

Pentru determinarea lui $g(x)$ trebuie să rezolvăm ecuația funcțională (II.7,6), adică

$$\Delta_a g(x) = m\left(x+\frac{a}{2}\right)^{m-1}.$$

Soluția principală a acestei ecuații este

$$g(x) = B_m\left(\frac{x}{a}+\frac{1}{2}\right),$$

unde $B_m(x)$ este polinomul lui Bernoulli de gradul m.
In baza relației (II.7,7) vom avea

$$\int_{-\infty}^{+\infty} x^m f(x)\,dx = \sum_{i=-\infty}^{+\infty} B_m\left(\frac{x_i}{a}+\frac{1}{2}\right) \int_{-\frac{a}{2}}^{\frac{a}{2}} f(x_i+u)\,du.$$

Știind că

$$B_m\left(\frac{x}{a}+\frac{1}{2}\right) = \sum_{i=0}^{m} C_m^i B_{m-i}\left(\frac{a}{2}\right)^{m-i} x^i.$$

B_{m-i} fiind numerele lui Bernoulli, rezultă

$$M_m=\sum_{i=0}^{m} C_m^i B_{m-i}\left(\frac{a}{2}\right)^{m-i} M'_m, \qquad \text{(II.7,8)}$$

deci formulele avute în vedere.

Particularizînd în relaţia (II.7,8) valorile lui m, obţinem valorile momentelor obişnuite M_m cu corecţiile Sheppard arătate în relaţia (II.7,1).

2) $$\varphi(x)=x(x-1)\ldots(x-m+1)=(x)_m.$$

In acest caz, soluţia principală a ecuaţiei (II.7,6) este

$$g(x)=\sum_{i=0}^{m} C_m^i (x_i)_{m-j} a^j \left(B_j\left(\frac{3}{2}\right)\right)_{j+2}.$$

Relaţia (II.7,7) devine

$$\int_{-\infty}^{+\infty}(x)_m f(x)\,dx=\sum_{i=0}^{m} C_m^j a^j \left(B_j\left(\frac{3}{2}\right)\right)_{j+2}(x_i)_{m-j}\int_{-\frac{a}{2}}^{\frac{a}{2}} f(x_i+u)\,du \quad (R=0).$$

Dezvoltînd, obţinem

$$M_m=\sum_{j=0}^{m} C_m^j a^j \left(B_j\left(\frac{3}{2}\right)\right)_{j+2} M'_{(m-j)}. \qquad \text{(II.7,9)}$$

De unde, particularizînd în această relaţie valorile lui m, găsim valorile momentelor factoriale $M_{(m)}$ din relaţia (II.7,2).

3) $$\varphi(x)=e^{tx}.$$

Avem ca soluţie principală a ecuaţiei (II.7,6)

$$g(x)=\frac{at}{e^{at}-1}e^{t\left(x+\frac{a}{2}\right)}$$

şi deci putem scrie în virtutea relaţiei fundamentale (II.7,7)

$$\int_{-\infty}^{+\infty} e^{tx} f(x)\,dx=\sum_{i=-\infty}^{+\infty}\frac{at}{e^{at}-1}e^{t\left(x_i+\frac{a}{2}\right)}\int_{-\frac{a}{2}}^{\frac{a}{2}} f(x+u)\,du,$$

în care restul, prin considerații analoge, se dovedește a fi de același ordin de mărime. Deci

$$\int_{-\infty}^{+\infty} e^{tx} f(x)\,dx = \sum_{i=-\infty}^{+\infty} \frac{at\,e^{\frac{at}{2}}}{e^{at}-1} \int_{-\frac{a}{2}}^{\frac{a}{2}} f(x+u)\,du = \frac{at\,e^{\frac{at}{2}}}{e^{at}-1} M'(e^{tx}).$$

In definitiv

$$M(e^{tx}) = \frac{at\,e^{\frac{at}{2}}}{e^{at}-1} \cdot M'(e^{tx}). \qquad \text{(II.7,10)}$$

Luînd derivata logaritmică a relației (II.7,10) în raport cu t avem

$$\frac{\partial \ln M(e^{tx})}{\partial t} = \frac{1}{t} + \frac{a}{2} - \frac{ae^{at}}{e^{at}-1} + \frac{\partial \ln M'(e^{tx})}{\partial t}.$$

Cum însă

$$M(e^{tx}) = e^{\frac{t}{1} M_{[1]} + \frac{t^2}{2!} M_{[2]} + \ldots + \frac{t^n}{n!} M_{[n]} + \ldots},$$

rezultă

$$\frac{\partial \ln M(e^{tx})}{\partial t} = M_{[1]} + \frac{t}{1} M_{[2]} + \ldots + \frac{t^{n-1}}{(n-1)!} M_{[n-1]} + \ldots,$$

$$\frac{\partial \ln M'(e^{tx})}{\partial t} = M'_{[1]} + \frac{t}{1} M'_{[2]} + \ldots + \frac{t^{n-1}}{(n-1)!} M'_{[n-1]} + \ldots$$

și deci în baza celor de mai sus

$$\sum_{i=1}^{\infty} [M_{[i]} - M'_{[i]}] \frac{t^{i-1}}{(i-1)!} = -\sum_{i=0}^{\infty} B_i a^i \frac{t^{i-1}}{i!}.$$

Identificînd coeficienții puterilor lui t de același ordin, deducem

$$M_{[i]} = M'_{[i]} - \frac{B_i a^i}{i}, \qquad \text{(II.7,11)}$$

unde B_i sînt numerele lui Bernoulli.
Deoarece numerele lui Bernoulli de ordin impar B_{2i+1} sînt nule începînd cu ordinul al treilea, avem

$$M_{[2i+1]} = M'_{[2i+1]} \qquad (i = 1, 2, \ldots).$$

In particular, pentru valorile lui $i = 1, 2, \ldots$, obținem valorile cumulanților $M_{[i]}$ din relația (II.7,3).

Scanned with OKEN Scanner

Capitolul III

Valori tipice uzuale

§ 1. Considerații generale

In capitolele precedente am văzut că Statistica dispune de diferite procedee cu ajutorul cărora prezintă datele furnizate de observații într-un mod adecvat scopului către care este îndreptată cercetarea. Prezentarea datelor observate sub forma unei tabele statistice este primul procedeu prin care statisticianul desprinde din ansamblul de observații, caracteristicile fenomenului studiat. Aceasta exprimă necesitatea pentru oricare cercetător de a reduce la un volum mic, cantitatea considerabilă de date observate. Numai în felul acesta poate cuprinde esențialul din mulțimea observațiilor asupra unuia sau mai multor fenomene.

Așa se justifică reprezentarea grafică a unei distribuții statistice, care scoate în evidență unele elemente specifice ale acesteia și dă, mai mult decît înfățișarea printr-o simplă tabelă, posibilitatea de a compara între ele diferite distribuții. Astfel, se pot compara regiunile unei valori centrale, care caracterizează concentrarea și dispunerea valorilor variabilei, precum și gradul de asimetrie al distribuției.

Este cîteodată necesar să exprimăm anume informații conținute în mulțimea de date observate, prin cîte o valoare numerică.

Asemenea *valori*, *caracteristice ale distribuției*, am întîlnit mai sus sub denumirea de *momente*, *funcție caracteristică* și *cumulanți* ai distribuției. Cunoașterea lor poate determina în mod unic distribuția; prin ele statistica dispune de un instrument de cercetare mai puternic decît simpla înfățișare a unei tabele sau reprezentarea sa grafică. In cele ce urmează examinăm acele caracteristici care prin rolul lor reprezentativ în studiul distribuțiilor sînt ceea ce se numește îndeobște *valori tipice*.

§ 2. Media aritmetică

Am definit media aritmetică a unei distribuții a cărei funcție de distribuție este $F(x)$, prin integrala (Stieltjes)

$$M = \int_{-\infty}^{+\infty} x\, dF(x), \qquad \text{(III.2,1)}$$

atunci cînd această valoare este finită.

Această expresie se reduce la integrala ordinară în cazul distribuției continue

$$M=\int_{-\infty}^{+\infty} x f(x)\,dx, \quad (dF(x)=f(x)\,dx, \qquad \text{(III.2,2)}$$

iar în cazul distribuției directe la suma

$$M=\sum_{k=-\infty}^{+\infty} x_k s_k, \qquad \text{(III.2,3)}$$

unde am notat ca de obicei prin s_k frecvența relativă a valorii x_k.
In aplicații se lucrează de multe ori cu frecvențele absolute și de aceea expresiile (III.2,2) și (III.2,3) ale mediei aritmetice se vor scrie în felul următor

$$M=\frac{1}{m}\int_{-\infty}^{+\infty} xg(x)\,dx, \qquad M=\frac{1}{n}\sum_{k=-\infty}^{+\infty} x_k u_k, \qquad \text{(III.2,4)}$$

unde acum $g(x)$ și u_k sînt respectiv frecvențele absolute, iar m și n, respectiv volumele distribuțiilor.
Expresia mediei aritmetice are unele proprietăți caracteristice. Mai jos arătăm unele dintre aceste proprietăți.
Valoarea medie a abaterilor unei variabile x față de media sa aritmetică este nulă.
Media aritmetică este singura valoare tipică ce se bucură de această proprietate.
Teoremă. *Media aritmetică a unei sume de variabile este egală cu suma mediilor aritmetice a fiecărei variabile în parte.*
Teoremă. *Valoarea minimă a expresiei*

$$\int_{-\infty}^{+\infty} (x-a)^2\,dF(x),$$

în caz că această expresie este finită, are loc pentru $a=M$.
In adevăr

$$\int_{-\infty}^{+\infty} [(x-M)+(M-a)]^2 dF(x)=\int_{-\infty}^{+\infty} [(x-M)^2+2(x-M)(M-a)+$$

$$+(M-a)^2]\,dF(x)=\int_{-\infty}^{+\infty} [(x-M)^2+(M-a)^2]\,dF(x).$$

de unde,

$$\int_{-\infty}^{+\infty}(x-a)^2\,dF(x) > \int_{-\infty}^{+\infty}(x-M)^2\,dF(x), \quad \text{dacă} \quad a \neq M.$$

Teorema este astfel demonstrată.
Calculul mediei aritmetice rezultă uşor, folosind alegerea unei origini arbitrare a, cît mai aproape de media aritmetică evaluată aproximativ, cînd aceasta se poate face, sau cît mai aproape de abscisa corespunzătoare frecvenţei maxime.
Procedînd astfel, obţinem

$$M = \int_{-\infty}^{+\infty}[(x-a)+a]\,dF(x) = a + M_a,$$

înţelegînd prin M_a media aritmetică faţă de noua origine în punctul de abscisă a.
Pentru calculul efectiv al mediei aritmetice, ne vom referi la valorile numerice ale momentelor arătate în exemplele 1 şi 2 (cap. II, § 8). Valoarea numerică a momentului de ordinul întîi coincide cu valoarea numerică a mediei aritmetice. După cum s-a arătat acolo, nu este nevoie de nici o corecţie pentru valoarea numerică a momentului de ordinul întîi şi deci, nici pentru aceea a mediei aritmetice. Aceasta este valabil, bineînţeles, numai pentru distribuţiile ale căror frecvenţe sînt egale cu zero la capetele intervalului de oscilaţie a distribuţiei (intervalul care cuprinde mulţimea valorilor variabilei).

§ 3. Mediana

Definiţie. *Mediana unei distribuţii este una dintre valorile* ξ *ale variabilei, pentru care avem*

$$\int_{-\infty}^{\xi} dF(x) = \int_{\xi}^{+\infty} dF(x) = \frac{1}{2}, \qquad \text{(III.3,1)}$$

$F(x)$ *fiind funcţia de distribuţie.*
In cazul unei distribuţii continue, mediana este unic determinată de egalitatea

$$\int_{-\infty}^{\xi} f(x)\,dx = \int_{\xi}^{+\infty} f(x)\,dx = \frac{1}{2}, \quad [dF(x) = f(x)\,dx]. \qquad \text{(III.3,2)}$$

In cazul distribuţiilor discrete, cînd relaţia (III.3,1) se va scrie

$$\sum_{k\leq \xi} s_k = \sum_{k\geq \xi} s_k = \frac{1}{2}, \qquad \text{(III.3,3)}$$

unde s_k reprezintă frecvenţele relative, este vizibil că există un interval întreg în care putem alege, în mod arbitrar, mediana.
Vom menţiona acum o proprietate deosebită a medianei, care o face preţioasă în aplicaţii.

Teoremă. *Suma abaterilor în raport cu un punct de abscisă a este minimă, cînd această abscisă coincide cu valoarea medianei.*
Luînd punctul de abscisă a şi un alt punct arbitrar de abscisă b $(a<b)$, considerăm următoarele cazuri

1) $x \leq a < b$.
Atunci

$$\int_{-\infty}^{a} |x-a|\, dF(x) = \int_{-\infty}^{a} (a-x)\, dF(x) = \int_{-\infty}^{a} (b-x)\, dF(x) -$$

$$-(b-a)\int_{-\infty}^{a} dF(x). \qquad \text{(III.3,4)}$$

2) $a \leq x \leq b$.
Avem

$$\int_{a}^{b} |x-a|\, dF(x) = \int_{a}^{b} (x-a)\, dF(x) = -\int_{a}^{b} (b-x)\, dF(x) +$$

$$+(b-a)\int_{a}^{b} dF(x). \qquad \text{(III.3,5)}$$

3) $a<b\leq x$.
In acest caz,

$$\int_{b}^{+\infty} |x-a|\, dF(x) = \int_{a}^{b} (x-a)\, dF(x) = \int_{b}^{+\infty} (x-b)\, dF(x) +$$

$$+(b-a)\int_{b}^{+\infty} dF(x). \qquad \text{(III.3,6)}$$

Adunînd membru cu membru (III.3,4), (III.3,5) şi (III.3,6), rezultă

$$\int_{-\infty}^{+\infty} |x-a|\,dF(x) = \int_{-\infty}^{+\infty} |x-b|\,dF(x) - 2\int_{a}^{b} (b-x)\,dF(x) -$$

$$-(b-a)\Big[\int_{-\infty}^{a} dF(x) - \int_{a}^{+\infty} dF(x)\Big]. \qquad \text{(III.3,7)}$$

Dacă luăm acum $a=\xi$ (valoarea medianei), egalitatea (III.3,7) devine

$$\int_{-\infty}^{+\infty} |x-\xi|\,dF(x) = \int_{-\infty}^{+\infty} |x-b|\,dF(x) - 2\int_{a}^{b} (b-x)\,dF(x), \qquad \text{(III.3,8)}$$

adică

$$\int_{-\infty}^{+\infty} |x-\xi|\,dF(x) < \int_{-\infty}^{+\infty} |x-b|\,dF(x). \qquad \text{(III.3,9)}$$

Inegalitatea (III.3,9) ne arată că suma abaterilor absolute ale variabilei în raport cu valoarea medianei este mai mică decît aceeaşi sumă a abaterilor, luate în raport cu abscisa oricărui alt punct.

In practică, pentru determinarea medianei este necesar să facem unele precizări.

Aşa de pildă, în cazul unei distribuţii discrete şi uniforme cu un volum de $2n+1$ elemente, mediana este reprezentată de elementul de rang $n+1$. Dacă distribuţia are un volum de $2n$ elemente, pentru mediană se ia media aritmetică a elementelor de rangul n şi $n+1$.

In ceea ce priveşte distribuţiile continue grupăm valorile pe intervale egale şi le repartizăm uniform în interiorul fiecărui interval care este reprezentat prin centrul său.

Exemplul 1. Valoarea medianei distribuţiei a 2 000 de serii în raport cu numărul numerelor prime (tabela I.2,2).

Deoarece volumul distribuţiei este egal cu 2 000, atunci considerăm elementele de rang 1 000 şi 1 001. Dăm mai jos, distribuţia completată cu coloana frecvenţelor cumulate (tabela III.3,1).

Din simpla exprimare a acestei distribuţii observăm că 1 192 serii au mai puţin de 33 numere prime fiecare şi 998, mai puţin de 32 numere fiecare. Atunci, rezultă că seriile de rang 1 000 şi 1 001 sînt cuprinse în grupa celor cu 32 numere prime fiecare şi deci, valoarea medianei este $\xi=32$.

Scanned with OKEN Scanner

TABELA III.3,1

Frecvența numerelor prime într-un număr de 2 000 serii a cîte 500 numere întregi consecutive fiecare, cuprinse între numerele 9 000 000 și 10 000 000

x	$f(x)$	$\Sigma f(x)$	x	$f(x)$	$\Sigma f(x)$	x	$f(x)$	$\Sigma f(x)$
18	1	1	27	130	305	36	89	1 706
19	1	2	28	135	440	37	72	1 878
20	7	9	29	171	611	38	54	1 932
21	9	18	30	177	788	39	27	1 959
22	13	31	31	210	998	40	20	1 979
23	20	51	32	194	1 192	41	11	1 990
24	46	97	33	171	1 363	42	4	1 994
25	73	170	34	143	1 506	43	4	1 998
26	105	275	35	111	1 617	44	2	2 000

Exemplul 2. Mediana distribuției precipitațiilor atmosferice în raport cu cantitatea de apă (tabela I.2,3).
Pentru o mai bună orientare, dăm la fel distribuția completată cu coloana frecvențelor cumulate (tabela III.3,2).

TABELA III.3,2.

Distribuția precipitațiilor atmosferice la noi în țară în intervalul 1896—1938, în timpul verii, în raport cu cantitatea de apă

x (mijlocul intervalului)	$f(x)$	$\Sigma f(x)$	x	$f(x)$	$\Sigma f(x)$
115	2	2	235	6	35
145	6	8	265	1	36
175	6	14	295	3	39
205	15	29	325	1	40

Jumătate din volumul acestei distribuții este 20, care corespunde unei valori din intervalul (175—205) de valori ale variabilei. In baza ipotezii repartiției uniforme a frecvențelor în fiecare interval, putem scrie pentru valoarea medianei

$$\xi = 175 + \frac{40 - 2 \times 14}{2 \times 15} \times 30 = 187.$$

In cazul cînd valoarea medianei se poate lua cu oarecare aproximație, putem folosi în acest scop reprezentarea grafică a funcției de distribuție $y = F(x)$.
Obținem așa-numita *ogivă a lui Galton*, luînd în sistemul coordonatelor rectangulare valorile argumentului ca abscise și ale frecvențelor cumulate ca ordonate.

Scanned with OKEN Scanner

Graficul ogivei se prezintă ca o scară cu trepte nedescrescătoare, în cazul unei distribuţii discrete (fig. III.3,1 *a*) şi ca o linie poligonală crescătoare, în cazul unei distribuţii continue cu valorile variabilei grupate pe intervale egale (fig. III.3,1 *b*). Ordonata ultimei abscise va corespunde bineînţeles frecvenţei cumulate, egală cu volumul distribuţiei. Ducînd o paralelă la axa absciselor prin mijlocul acestei ordonate, tăiem ogiva într-un punct a cărui abscisă este mediana căutată.

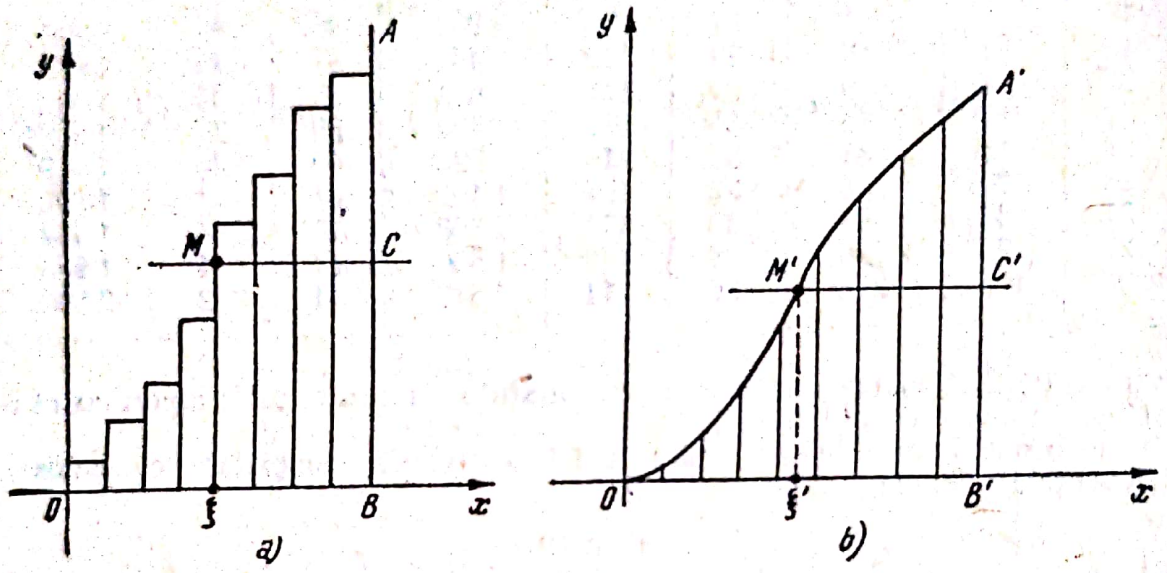


Fig. III.3,1

Din studiul medianei se desprind unele observaţii interesante de care statisticianul trebuie să ţină seama în rezolvarea problemelor practice.

In primul rînd, mediana este o valoare tipică mult mai independentă decît media aritmetică, faţă de scara şi forma distribuţiei. In privinţa aceasta mediana este extrem de folositoare, mai ales pentru distribuţiile deschise, unde limitele intervalului care conţine mulţimea valorilor variabilei nu sînt fixate. In cazul unor asemenea distribuţii, media aritmetică conduce la valori mai puţin concludente.

In al doilea rînd, mediana fiind mai puţin influenţată de fiecare dintre elementele distribuţiei, poate fi utilizată cu mai multă încredere în cercetările unde influenţa elementelor străine nu a putut fi înlăturată.

Mediana prezintă însă şi unele inconveniente, ca de pildă, lipsa teoremei de adiţiune, care este valabilă pentru media aritmetică. Prin urmare, în cazul contopirii mai multor distribuţii, va trebui să calculăm din nou mediana distribuţiei obţinute, neputînd altfel să avem nici o indicaţie pe baza valorii medianei fiecărei distribuţii în parte.

Scanned with OKEN Scanner

§ 4. Quantile

Din consideraţiile făcute mai înainte, am văzut că mediana, în unele împrejurări, ne poate da indicaţii mai bune despre o distribuţie, decît media. Utilizînd aceste calităţi, putem extinde noţiunea de mediană în vederea obţinerii unor mai mari posibilităţi în studiul distribuţiilor.

Astfel, ne propunem a găsi valorile ξ_i $(i=1,\ 2,\ \ldots,\ n-1)$ ale variabilei, pentru care avem

$$\int_{-\infty}^{\xi_1} dF(x)=\int_{\xi_1}^{\xi_2} dF(x)=\ldots=\int_{\xi_{n-1}}^{+\infty} dF(x)=\frac{1}{n}, \qquad \text{(III.4,1}$$

$F(x)$ fiind funcţia de distribuţie.

Valorile ξ_i se numesc în general *quantile* şi, după cum rezultă din egalităţile (III.4,1), ele împart volumul S al distribuţiei în n părţi egale. In aplicaţii, statistica întrebuinţează unele valori particulare ale quantilelor. Dacă $n=4$, de exemplu, avem trei valori $(\xi_1,\ \xi_2,\ \xi_3)$ numite *quartile,* care împart volumul distribuţiei în patru părţi egale. Este evident că a doua quantilă, ξ_2, coincide cu mediana ξ a distribuţiei. Raportate la valoarea medianei, quantilele ξ_1 şi ξ_2 se numesc respectiv *quartila mică (inferioară)* şi *quartila mare (superioară).*

Un alt caz particular al quantilelor rezultă luînd $n=10$. Cele nouă valori $\xi_1, \xi_2, \ldots, \xi_9$, numite *decile*, împart volumul distribuţiei în zece părţi egale. Decilele dau o măsură mai bună decît quartilele în ceea ce priveşte forma distribuţiei cercetate.

Pentru calculul efectiv al quantilelor în general şi al decilelor şi quartilelor în special, procedăm la fel ca la calculul medianei, rămînînd valabile şi aici precizările făcute acolo, atît pentru distribuţiile discrete, cît şi pentru cele continue.

Exemplul 1. Referindu-ne la distribuţia celor 2 000 serii în raport cu numărul numerelor prime (tabela I.2,2) care are mediana $\xi=32$, să determinăm quartilele şi decilele.

Găsim următoarele valori ale quartilelor, ţinînd seama că 1/4 din volumul distribuţiei este egal cu 500,

$$\xi_1=29,\quad \xi_2=\xi=32,\quad \xi_3=34.$$

Pentru decile obţinem, observînd că 1/10 din volumul distribuţiei este egal cu 200

$$\xi_1=26,\quad \xi_2=28,\quad \xi_3=29,\ \ldots,\quad \xi_8=35,\quad \xi_9=37.$$

Exemplul 2. In cazul distribuţiei continue a precipitaţiilor atmosferice în raport cu cantitatea de apă (tabela I.2,3), rezultă următoarele valori pentru quartile

$$\xi_1 = 145 + \frac{40-4\cdot 8}{2\cdot 6}\cdot 30 = 165,$$

$$\xi_2 = 175 + \frac{40-2\cdot 14}{2\cdot 15}\cdot 30 = 187 \text{ (mediana)},$$

$$\xi_3 = 205 + \frac{40-29}{2\cdot 6}\cdot 30 = 232{,}5.$$

In mod analog, se pot obţine şi valorile celor nouă decile $\xi_1, \xi_2, \ldots, \xi_9$.

Se poate întrebuinţa pentru calculul quantilelor, în general, ogiva lui Galton, despre care am vorbit în paragraful precedent, dacă nu este nevoie de o prea mare precizie a lor. Astfel, quantila de rangul k va fi abscisa punctului de intersecţie a ogivei cu paralela la axa x-ilor, dusă prin diviziunea k a ordonatei corespunzătoare frecvenţei cumulate, egală cu volumul distribuţiei. Numărarea diviziunilor (egale între ele) se consideră pe ordonată, de la axa x-ilor.

§ 5. Modul

Definiţie. *Modul unei distribuţii este*, prin definiţie, *acea valoare a variabilei căreia îi corespunde frecvenţa maximă.*

Modul este o noţiune legată de distribuţiile continue şi este, cu alte cuvinte, abscisa vîrfului curbei de distribuţie, dacă această curbă are un singur vîrf determinat.

In aceste condiţii, valoarea modului este soluţia ecuaţiei

$$\frac{df(x)}{dx} = 0 \qquad \left(\frac{d^2 f(x)}{dx^2} < 0\right), \tag{III.5,1}$$

unde presupunem că funcţia de frecvenţă $f(x)$ este continuă, împreună cu primele două derivate.

Un minim al curbei distribuţiei poartă numele de *antimod*. Antimodul este, deci, în particular, o valoare tipică a distribuţiilor în formă de U.

In practică se întrebuinţează diferite procedee care dau cu aproximaţie valoarea modului.

Ca o primă aproximaţie, se consideră ca valoare a modului valoarea variabilei din mijlocul intervalului corespunzător frecvenţei maxime.

Scanned with OKEN Scanner

Un procedeu, care permite aflarea modulului cu o aproximaţie satisfăcătoare, constă în aplicarea formulei

$$x_k = \frac{f_{k-1} - f_{k-1}}{2f_k - f_{k+1} - f_{k-1}} \cdot \frac{d}{2}, \qquad \text{(III.5,2)}$$

unde x_k este mijlocul intervalului de frecvenţă maximă, f_k, f_{k-1} şi f_{k+1} respectiv frecvenţa maximă şi frecvenţele vecine ei, iar d lungimea intervalului în care sînt grupate valorile variabilei.

Arătăm pe scurt cum se determină formula aproximativă (III.5,2).

Să considerăm intervalul de frecvenţă maximă al unei distribuţii continue, împreună cu intervalele de frecvenţă vecine lui

$$\left(x_{k-1} - \frac{1}{2},\ x_{k-1} + \frac{1}{2}\right),\quad \left(x_k - \frac{1}{2},\ x_k + \frac{1}{2}\right),$$

$$\left(x_{k+1} - \frac{1}{2},\ x_{k+1} + \frac{1}{2}\right)$$

cărora le corespund frecvenţele

$$f_{k-1} < f_k < f_{k+1}.$$

Ajustăm distribuţia în intervalul $\left(x_{k-1} - \frac{1}{2},\ x_{k+1} + \frac{1}{2}\right)$ cu ajutorul arcului de parabolă $y = a + bx + cx^2$, ai cărei coeficienţi a, b şi c se pot determina din ecuaţiile

$$\int_{x_{k-1}-\frac{1}{2}}^{x_{k-1}+\frac{1}{2}} y\,dx = f_{k-1}, \qquad \int_{x_k-\frac{1}{2}}^{x_k+\frac{1}{2}} y\,dx = f_k, \qquad \int_{x_{k+1}-\frac{1}{2}}^{x_{k+1}+\frac{1}{2}} y\,dx = f_{k+1}. \qquad \text{(III.5,3)}$$

Modul va fi dat de abscisa vîrfului parabolei, a cărei valoare se obţine din ecuaţia

$$\frac{dy}{dx} = b + 2cx = 0. \qquad \text{(III.5,4)}$$

Inlocuind acum în ecuaţia (III.5,4), valorile coeficienţilor a, b şi c, date de ecuaţiile (III.5,3), deducem formula (III.5,2).

Scanned with OKEN Scanner

Pentru distribuţiile uşor asimetrice, K. Pearson a stabilit următoarea formulă simplă, pentru calculul modului

$$M-3(M-\xi), \qquad \text{(III.5,5)}$$

unde am notat prin M şi ξ, respectiv media aritmetică şi mediana. In ceea ce priveşte distribuţiile discrete, se citeşte direct valoarea argumentului corespunzător frecvenţei maxime.

De exemplu, pentru distribuţia discretă a celor 2 000 serii în raport cu numărul numerelor prime (tabela I.2,2), modul este egal cu 31, deoarece aceasta este valoarea argumentului corespunzătoare frecvenţei maxime egală cu 210.

Pentru distribuţia continuă a precipitaţiilor atmosferice în raport cu cantitatea de apă (tabela I.2,3), putem aplica formula simplă (III.5,5) întrucît această distribuţie îndeplineşte condiţia de asimetrie uşoară. Găsim astfel, valoarea modului egală cu

$$202-3(202-187)=157.$$

Există, după cum am arătat mai înainte (cap. I, § 4), distribuţii cu mai multe vîrfuri, numite *multimodale*, care au un *mod principal* şi mai multe *secundare*.

In special, modul ca valoare tipică a unei distribuţii se foloseşte atunci, cînd ne interesează cea mai frecventă valoare, ca de exemplu orele de chemări telefonice mai numeroase, de aglomeraţie a călătorilor etc., cu scopul unor mai bune măsuri de ordin organizatoric.

Aceleaşi observaţii făcute cu privire la mediană rămîn valabile şi pentru mod.

In cazul unei distribuţii simetrice, în mod teoretic media, mediana şi modul coincid, ceea ce contribuie ca acest fapt să fie un procedeu bun de caracterizare a unei astfel de distribuţii.

§ 6. Extinderea sau intervalul de variaţie

Definiţie. *Extinderea sau intervalul de variaţie al distribuţiei* se numeşte *diferenţa dintre marginea superioară şi cea inferioară a variabilei.*

Dacă α şi β reprezintă respectiv marginea inferioară şi cea superioară a variabilei, atunci intervalul de variaţie al distribuţiei este dat de diferenţa

$$\beta-\alpha. \qquad \text{(III.6,1)}$$

Intervalul de variaţie (III.6,1), nu poate fi luat ca o măsură suficientă a *dispersiei*, adică a modului de repartizare a valorilor variabilei în jurul uneia dintre valorile tipice, media aritmetică, mediana sau modul. Cît priveşte distribuţiile cu intervalele de variaţie infinite, este de la sine înţeles că acestea din urmă rămîn fără obiect.

Evaluări ale dispersiei, mai des folosite decît intervalul de variaţie, sînt date de *intervalele interquantile*, reprezentate de distanţa $(\xi_{n-1}-\xi_1)$ dintre ultima quantilă, de rangul $n-1$ şi prima.

În aplicaţii se iau de obicei intervalele *interdecile* şi *interquartile*. Intervalul interdecil este distanţa dintre decila a noua şi prima $(\xi_9-\xi_1)$, iar cel interquartil, distanţa dintre quartilul superior şi inferior $(\xi_3-\xi_1)$.

În cazul distribuţiilor simetrice, avem egalitatea

$$\xi-\xi_1=\xi_3-\xi, \qquad \text{(III.6,2)}$$

unde cu notaţiile obişnuite ξ, ξ_1 şi ξ_3 sînt respectiv mediana şi cele două quartile, inferior şi superior.

Egalitatea (III.6,2) ne spune că intervalele interquartile $\xi-\xi_1$ şi $\xi_3-\xi$ reprezintă de asemenea indicaţii bune pentru dispersia distribuţiilor simetrice. Dar această egalitate nu mai are loc pentru distribuţii uşor sau pronunţat asimetrice şi de aceea se adoptă ca măsură a dispersiei, media aritmetică

$$\frac{(\xi-\xi_1)+(\xi_3-\xi)}{2}=\frac{\xi_3-\xi_1}{2}. \qquad \text{(III.6,3)}$$

§ 7. Media abaterilor absolute

Definiţie. *Media abaterilor absolute ale variabilei unei distribuţii este, prin definiţie*

$$\int_{-\infty}^{+\infty}|x-a|\,dF(x), \qquad \text{(III.7,1)}$$

unde a este o valoare arbitrară din mulţimea valorilor variabilei, cuprinse în intervalul de variaţie al distribuţiei, *dacă această integrală există.*

In general, media abaterilor absolute se calculează în raport cu media aritmetică sau mediana distribuţiei, în care caz expresia (III.7,1) devine

$$\int_{-\infty}^{+\infty} |x-M|\,dF(x) \quad \text{sau} \quad \int_{-\infty}^{+\infty} |x-\xi|\,dF(x), \qquad \text{(III.7,2)}$$

M şi ξ fiind, conform notaţiilor noastre, respectiv media aritmetică şi mediana.

Vom menţiona, de asemenea, că expresiile (III.7,2) se vor scrie cu ajutorul integralei ordinare

$$\int_{-\infty}^{+\infty} |x-M|\,f(x)\,dx \quad \text{sau} \quad \int_{-\infty}^{+\infty} |x-\xi|\,f(x)\,dx$$

dacă distribuţia are o funcţie de frecvenţă continuă şi cu suma obişnuită

$$\sum_{k=-\infty}^{+\infty} |x_k-M|\,s_k \quad \text{sau} \quad \sum_{k=-\infty}^{+\infty} |x_k-\xi|\,s_k,$$

cînd distribuţia este discretă, s_k fiind frecvenţele relative.

La noţiunea de medie a abaterilor absolute sîntem conduşi de necesitatea de a cunoaşte cît mai multe aspecte ale distribuţiei. Valorile tipice ca media aritmetică, mediana şi modul nu ne dau decît un anumit aspect al distribuţiei şi anume, ordinul de mărime mijlociu al frecvenţelor sale.

Pentru cercetarea altor aspecte ale distribuţiei, ca de pildă, modul de repartizare a frecvenţelor sale în jurul unei valori medii, ceea ce numim împrăştiere, este nevoie de a cunoaşte, în afară de intervalul de variaţie, unele intervale interquantile şi anume momente sau valori medii raportate la una din valorile tipice.

Exemplu. Să considerăm următoarele distribuţii, care reprezintă numărul elevilor din două clase în raport cu notele obţinute la învăţătură

$$X_1 \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 0 & 10 & 21 & 16 & 3 \end{pmatrix},$$

$$X_2 \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 10 & 5 & 11 & 11 & 13 \end{pmatrix}.$$

Calculînd media aritmetică a fiecărei distribuţii, avem

$$M(X_1)=M(X_2)=3{,}24.$$

Deci cele două distribuţii au aceeaşi medie aritmetică.

Acest rezultat ne confirmă cele spuse mai sus că media aritmetică, ca valoare tipică, nu ne oferă nici o indicaţie asupra împrăştierii elementelor distribuţiilor.

Un simplu grafic al distribuţiilor, folosind poligonul frecvenţelor, ne este sugestiv (fig. III.7,1).

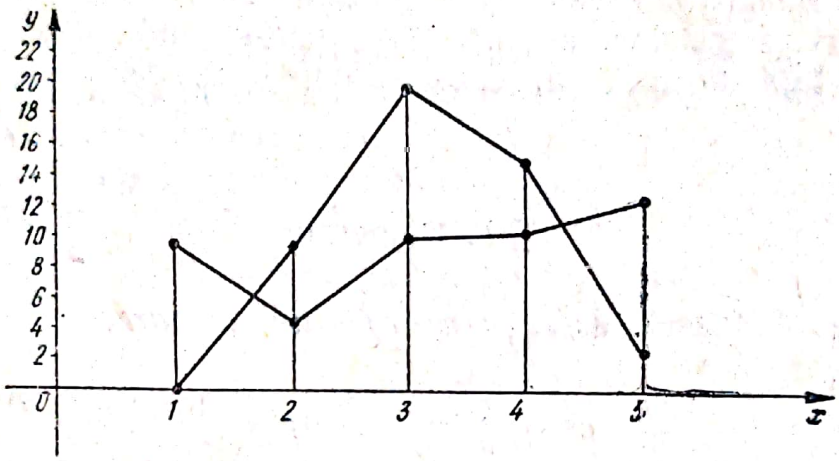


Fig. III.7,1

Valoarea numerică a mediei abaterilor absolute, în raport cu media aritmetică calculată este pentru fiecare dintre distribuţii (tabela III.7,1)

$$\frac{\sum_{k=1}^{5} |x_k-3{,}24|\,u_k}{\sum u_k}=0{,}697\,6\,; \qquad \frac{\sum_{k=1}^{5} |x_k-3{,}24|\,v_k}{\sum v_k}=1{,}249\,6.$$

TABELA III.7.1

x	Frecvenţa absolută u_k	$\lvert x_k-3{,}24\rvert$	$\lvert x_k-3{,}24\rvert u_k$	x	Frecvenţa absolută v_k	$\lvert x_k-3{,}24\rvert$	$\lvert x_k-3{,}24\rvert v_k$
1	0	2,24	0	1	10	2,24	22,40
2	10	1,24	12,40	2	5	1,24	6,20
3	21	0,24	5,04	3	11	0,24	2,64
4	16	0,76	12,16	4	11	0,76	8,36
5	3	1,76	5,28	5	13	1,76	22,88
	50		34,88		50		62,48

Cele două valori numerice obținute pentru mediile abaterilor absolute scot în evidență faptul că există împrăștieri diferite pentru frecvențele distribuțiilor considerate, ceea ce nu putem cunoaște numai pe baza mediei aritmetice.

In cele două exemple, am calculat media abaterilor absolute în raport cu media aritmetică.

Folosind proprietatea medianei cu privire la suma abaterilor absolute, care este minimă cînd este luată în raport cu mediana, putem aduce simplificări calculelor de mai sus, considerind media abaterilor absolute față de mediană, dacă valoarea acesteia se poate obține mai ușor.

§ 8. Dispersia

Definiție. *Dispersia este, prin definiție, valoarea*

$$\sigma^2 = \int_{-\infty}^{+\infty} (x-M)^2 \, dF(x), \qquad \text{(III.8,1)}$$

dacă membrul al doilea este finit.

Dispersia este, prin urmare, momentul centrat de ordinul al doilea al distribuției ($\sigma^2 = m_2$). Ea se mai notează și prin $D^2(x)$.

Valoarea

$$\sigma = \sqrt{m_2} \qquad \text{(III.8,2)}$$

se numește *abatere medie pătratică* și, după cum se vede, ea este o mărime de același ordin cu variabila x.

Dispersia și abaterea medie pătratică îndeplinesc aceeași funcție ca și media abaterilor absolute: de a fi valori tipice pentru împrăștiere, pe care o pun, însă, mai puternic în evidență.

De asemenea, momentele centrate de ordin par mai mare decît doi, precum și momentele centrate absolute de orice ordin, sînt adesea luate drept valori tipice ale împrăștierii. Vom folosi însă, cu deosebire, dispersia sau abaterea medie pătratică, pentru motivul că sînt cele mai simple dintre toate.

Vom deduce acum unele proprietăți ale dispersiei, de care ne vom servi la calcularea ei.

Scanned with OKEN Scanner

Teoremă. *Pentru orice valoare a, avem*

$$\sigma^2(x)=M_2(x-a)-(M_1-a)^2.$$

Intr-adevăr

$$M_2(x-a)=\int_{-\infty}^{+\infty}(x-a)^2\,dF(x)=\int_{-\infty}^{+\infty}[(x-M)+(M-a)]^2\,dF(x)=$$

$$=\sigma^2(x)+(M-a)^2.$$

Cu alte cuvinte, dispersia $\sigma^2(x)$ poate fi dedusă din momentul de ordinul al doilea în raport cu o valoare arbitrară a, aleasă astfel, ca să simplificăm calculele.

Teoremă. *Valoarea minimă a expresiei*

$$\int_{-\infty}^{+\infty}(x-a)^2\,dF(x)$$

este σ^2.

Din teorema precedentă, rezultă imediat

$$\int_{-\infty}^{+\infty}(x-a)^2\,dF(x)=M_2(x-a)>\sigma^2, \quad \text{dacă} \quad M_1\neq a,$$

ceea ce demonstrează teorema.

Teoremă. *Dispersia este independentă față de schimbarea originii.*

Am demonstrat că momentele centrate de orice ordin sînt independente față de originea aleasă (cap. II, § 3). Dispersia fiind momentul centrat de ordinul al doilea, teorema este evidentă.

Teoremă. *Pentru orice transformare liniară de forma* $x=ay+b$, *avem*

$$\sigma^2(x)=a^2\sigma^2(y).$$

Plecăm de la egalitatea

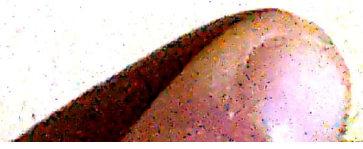
$$\sigma^2(x)=M_2(x)-M^2(x).$$

Făcînd schimbarea de variabilă $x=ay+b$, deducem

$$M_2(ay+b)-[M(ay+b)]^2=M(ay+b)-[aM(y)+b]^2=$$

$$=a^2[M_2(y)-M^2(y)]=a^2\sigma^2(y),$$

şi teorema este demonstrată.



Teoremă. *Dispersia unei sume de distribuţii este egală cu suma dispersiilor fiecărei distribuţii în parte, adunată cu suma pătratelor abaterilor mediilor aritmetice ale distribuţiilor date, în raport cu media aritmetică a sumei distribuţiilor.*

Prin definiţie, putem scrie

$$\sigma^2=\int_{-\infty}^{+\infty}(x-M)^2\,dF(x)=\sum_{i=1}^{n}\int_{-\infty}^{+\infty}(x-M)^2\,dF_i(x)=$$

$$=\sum_{i=1}^{n}(M_2^{(i)}-2M^{(i)}\cdot M+M^2)=\sum_{i=1}^{n}[M_2^{(i)}-(M^{(i)})^2+$$

$$+(M^{(i)})^2-2M^{(i)}\cdot M+M^2]=\sum_{i=1}^{n}\sigma_i^2+\sum_{i=1}^{n}(M^{(i)}-M)^2.$$

Teorema este deci demonstrată.

Consecinţă. Dispersia unei sume de distribuţie este mai mare sau egală cu suma dispersiilor individuale. Egalitatea are loc dacă şi numai dacă mediile distribuţiilor sînt egale între ele.

Din teoremă, rezultă

$$\sigma^2>\sum_{i=1}^{n}\sigma_i^2, \quad \text{dacă} \quad M^{(i)}\neq M;$$

$$(i=1,\ 2,\ \ldots,\ n).$$

$$\sigma^2=\sum_{i=1}^{n}\sigma_i^2, \quad \text{dacă} \quad M^{(i)}=M.$$

Exemplul 1. Dispersia celor două distribuţii ale numărului de elevi, în raport cu notele obţinute.

Orînduim calculele ca în tabela III.8,1.

$$\sigma_1^2=\frac{\sum_{k=1}^{5}(x_k-3{,}24)^2u_k}{\sum_{k=1}^{5}u_k}=\frac{35{,}120\,0}{50}=0{,}702\,4,$$

$$\sigma_2^2=\frac{\sum_{k=1}^{5}(x_k-3{,}24)^2v_k}{\sum_{k=1}^{5}v_k}=\frac{105{,}120\,0}{50}=2{,}102\,4.$$

TABELA III.8,1

x	u_k	$x_k-3,24$	$(x_k-3,24)^2$	$(x_k-3,24)^2u_k$	x	v_k	$x_k-3,24$	$(x_k-3,24)^2$	$(x_k-3,24)^2v_k$
1	0	−2,24	5,017 6	0	1	10	−2,24	5,017 6	50,176 0
2	10	−1,24	1,537 6	15,376 0	2	5	−1,24	1,537 6	7,688 0
3	21	−0,24	0,057 6	1,209 6	3	11	−0,24	0,057 6	0,633 6
4	16	0,76	0,577 6	9,241 6	4	11	0,76	0,577 6	6,353 6
5	3	1,76	3,097 6	9,292 8	5	13	1,76	3,097 6	40,268 8
	50			35,120 0		50			105,120 0

Observăm, astfel, că rapoartele valorilor numerice ale dispersiilor pentru cele două distribuţii sînt mai mari decît rapoartele valorilor numerice ale mediilor abaterilor absolute, calculate pentru aceleaşi distribuţii. Aceasta este evident consecinţa faptului că abaterile intervin în calcule prin pătratul lor, ceea ce contribuie ca abaterile mari (cazurile excepţionale) să influenţeze mai mult asupra valorii numerice a dispersiei şi, prin urmare, să pună mai puternic în evidenţă împrăştierea.

Exemplul 2. Dispersia se poate obţine mai uşor în urma unui calcul simplificat.

In acest scop, să considerăm distribuţia planetelor în raport cu valorile paralaxei (tabela I.2,4).

Calculele vor fi dispuse în modul următor (tabela III.8,2):

$$\sigma^2(y)=M_2(y)-[M(y)]^2=\frac{\sum y^2 g(x)}{\sum g(x)}-\left(\frac{\sum y\, g(x)}{\sum g(x)}\right)^2=$$

$$=\frac{31\,332}{640}-\left(\frac{2\,980}{640}\right)^2=48,95-(4,6)^2=27,79.$$

In baza transformării liniare

$$y=\frac{x-2,5}{5} \quad \text{sau} \quad x=5y+2,5,$$

putem trece imediat la valoarea dispersiei $\sigma^2(x)$

$$\sigma^2(x)=5^2\cdot\sigma^2(y)=25\cdot 27,79=694,75.$$

TABELA III.8.2

Mijlocul intervalului x	Frecvența absolută $g(x)$	$x-2,5$	$y=\frac{x-2,5}{5}$	y^2	$y \cdot g(x)$	$y^2 \cdot g(x)$	Mijlocul intervalului x	Frecvența absolută $g(x)$	$x-2,5$	$y=\frac{x-2,5}{5}$	y^2	$y \cdot g(x)$	$y^2 \cdot g(x)$
−27,5	3	−30	−6	36	−18	108	37,5	32	35	7	49	224	1 568
−22,5	7	−25	−5	25	−35	175	42,5	33	40	8	64	264	2 112
−17,5	7	−20	−4	16	−28	102	47,5	29	45	9	81	261	2 349
−12,5	17	−15	−3	9	−51	153	52,5	21	50	10	100	210	2 100
− 7,5	27	−10	−2	4	−54	108	57,5	12	55	11	121	132	1 452
− 2,5	39	− 5	−1	1	−39	39	62,5	17	60	12	144	204	2 448
2,5	46	0	0	0	0	0	67,5	13	65	13	169	169	2 197
7,5	51	5	1	1	51	51	72,5	15	70	14	196	210	2 940
12,5	56	10	2	4	112	224	77,5	12	75	15	225	180	2 700
17,5	50	15	3	9	150	450	82,5	6	80	16	256	96	1 536
22,5	43	20	4	16	172	638	87,5	6	85	17	289	102	1 734
27,5	45	25	5	25	225	1 125	92,5	5	90	18	324	90	1 620
32,5	43	30	6	36	258	1 548	97,5	5	95	19	361	95	1 805
								640				2 980	31 332

Aplicînd corecțiile Sheppard, deoarece am calculat dispersia unei distribuții continue cu valorile variabilei grupate pe intervale egale, putem scrie, ca valoare definitivă,

$$\sigma^2(x)=694,75-\frac{5^2}{12}=692,67.$$

Comparînd calculele care ne-au condus la valoarea numerică a mediei abaterilor absolute și dispersiei, observăm că cele pentru determinarea mediei abaterilor absolute sînt mai simple. În schimb, acest avantaj este urmat de unele dificultăți pe care nu le are dispersia.

În primul rînd, dispersia reprezintă o măsură mai precisă a împrăștierii, decît media abaterilor absolute.

În al doilea rînd, dispersia nu prezintă defecte de natură analitică, ca media abaterilor absolute, a cărei derivată în abscisele corespunzătoare frecvențelor este discontinuă. Din acest motiv, ea nu poate fi folosită în problemele teoretice mai avansate, ca de exemplu în teoria selecției.

Scanned with OKEN Scanner

In al treilea rînd, contrar dispersiei, media abaterilor absolute nu admite o teoremă de adiţiune, ceea ce îngreunează calculele cînd este vorba de contopirea mai multor distribuţii.

§ 9. Media diferenţelor absolute

Definiţie. *Media diferenţelor absolute ale valorilor variabilei unei distribuţii este valoarea*

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} |x-y|\, dF(x)\, dF(y). \qquad \text{(III.9,1)}$$

Din definiţia ei, rezultă că media diferenţelor absolute este media tuturor diferenţelor absolute posibile ale perechilor de valori ale variabilei.

Expresia (III.9,1) se va scrie evident

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} |x-y|\, f(x)\, f(y)\, dx\, dy$$

în cazul continuu şi

$$\frac{1}{s^2}\sum_{i=-\infty}^{+\infty}\sum_{j=-\infty}^{+\infty} |x_i-x_j|\, s_i s_j \quad \text{sau} \quad \frac{1}{s(s-1)}\sum_{i=-\infty}^{+\infty}\sum_{j=-\infty}^{+\infty} |x_i-x_j|\, s_i s_j$$

în cazul discontinuu, respectiv pentru media diferenţelor absolute cu repetiţie sau fără repetiţie a aceleiaşi valori a variabilei ($i \neq j$, $i=j$ sau $i \neq j$).

Media diferenţelor absolute a fost introdusă prima oară de Gini (1912) şi are mai mult o însemnătate teoretică, decît una practică. După cum se observă, ea este independentă de orice valoare tipică şi depinde numai de mulţimea valorilor variabilei.

La fel ca şi media abaterilor absolute, prezintă inconveniente de natură analitică, care o fac să fie neîntrebuinţată în teoriile statistice mai avansate, deosebit de faptul că se calculează mai greu decît dispersia.

Media pătratelor diferenţelor se bucură de o proprietate interesantă.

Teoremă. *Media pătratelor tuturor diferenţelor posibile ale perechilor de valori ale variabilei unei distribuţii este egală cu dublul dispersiei.*

Scanned with OKEN Scanner

Vom scrie

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(x-y)^2\,dF(x)\,dF(y)=\int_{-\infty}^{+\infty}x^2\,dF(x)\int_{-\infty}^{+\infty}dF(y)-$$

$$-2\int_{-\infty}^{+\infty}x\,dF(x)\int_{-\infty}^{+\infty}y\,dF(y)+\int_{-\infty}^{+\infty}dF(x)\int_{-\infty}^{+\infty}y^2\,dF(y)=2(M_2-M^2)=2\sigma^2.$$

Această teoremă poate fi considerată ca o altă definiţie a dispersiei.

§ 10. Coeficienţii de variabilitate şi concentraţie

Măsurile examinate pînă acum prezintă un dezavantaj, deoarece nu ne dau posibilitatea de a compara gradul de împrăştiere a două distribuţii ale căror variabile sînt exprimate în unităţi de măsură diferite. Aceasta se explică prin faptul că toate aceste măsuri pentru determinarea împrăştierii se exprimă în raport cu unităţile variabilei şi, prin urmare, valoarea lor numerică depinde de ordinul de mărime al elementelor distribuţiei. Este necesar, deci, să se caute pentru înlăturarea acestui neajuns alte măsuri, care să fie independente de felul unităţilor în care este exprimată variabila.
Asemenea măsuri pot fi obţinute în numere relative, care în terminologia statistică poartă numele de *coeficienţi*.
Dintre cei mai folosiţi în practică sînt *coeficientul de variabilitate* al lui Pearson, definit de expresia

$$100\,\frac{\sigma}{M} \qquad \text{(III.10,1)}$$

şi *coeficientul de concentraţie* al lui Gini, definit de expresia

$$\frac{\delta}{2M}, \qquad \text{(III.10,2)}$$

unde am notat cu δ, media diferenţelor absolute dată de (III.9,1).
Coeficientul de variabilitate al lui Pearson, definit de expresia (III.10,1) nu este altceva decît abaterea medie pătratică a unei distribuţii a cărei variabilă este $y=\frac{x}{M}$.

Intr-adevăr, avem

$$\sigma^2(y)=\frac{\sigma^2(x)}{M^2},$$

în baza transformării liniare $y=\frac{x}{M}$, de unde

$$100\,\sigma(y)=100\frac{\sigma(x)}{M}.$$

La coeficientul lui Gini se ajunge în mod natural, considerînd egalitățile

$$F(x)=\int_{-\infty}^{x} f(x)\,dx,$$

$$\Phi(x)=\frac{1}{M}\int_{-\infty}^{x} x f(x)\,dx. \qquad \text{(III.10,3)}$$

Funcția $\Phi(x)$ există, dacă și M există. Observăm că pentru valorile variabilei x cuprinse între $-\infty$ și $+\infty$, $F(x)$ și $\Phi(x)$ variază între 0 și 1.

Presupunem că originea este așezată la stînga intervalului de variație al distribuției.

Egalitățile (III.10,3) vor reprezenta ecuațiile parametrice ale unei curbe, numite *curba de concentrație,* luînd $F(x)$ ca abscisă și $\Phi(x)$ ca ordonată într-un sistem de axe rectangulare. Curba de concentrație este curbă convexă față de axa absciselor F, deoarece

$$\frac{d\Phi}{dF}=\frac{x}{M}>0, \quad \text{iar} \quad \frac{d^2\Phi}{dF^2}=\frac{1}{M\cdot f(x)}>0,$$

deoarece $x>0$, $f(x)>0$.

Rezultă că unghiul format de tangenta la curbă cu direcția pozitivă a axei F-ilor este ascuțit și crește o dată cu F.

Graficul curbei de concentrație este prezentat în fig. III.10,1.

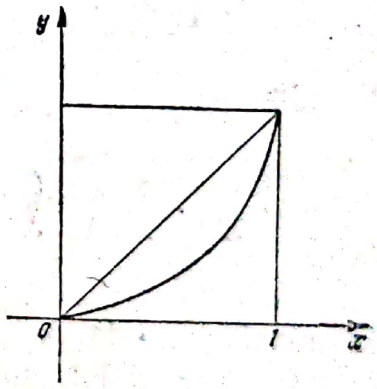


Fig. III. 10,1

Curba, împreună cu bisectoarea $\Phi=F$, mărginesc *aria de concentrație* Γ, pe care o vom determina imediat.

Observăm că

$$2\Gamma=\int_0^1 F\,d\Phi-\int_0^1 \Phi\,dF,$$

sau

$$2M\cdot\Gamma=\int_{-\infty}^{+\infty}F(x)\,x\,dx-M\int_{-\infty}^{+\infty}\Phi(x)\,dF(x)=$$

$$=\int_{-\infty}^{+\infty}x\,dF(x)\int_{-\infty}^{x}dF(y)-\int_{-\infty}^{+\infty}dF(x)\int_{-\infty}^{x}y\,dF(y)$$

şi deci

$$2M\cdot\Gamma=\int_{-\infty}^{+\infty}\int_{-\infty}^{x}(x-y)\,dF(x)\,dF(y).$$

Observînd, de asemenea, că

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(x-y)\,dF(x)\,dF(y)=0,$$

putem scrie

$$2M\Gamma=\frac{1}{2}\left[\int_{-\infty}^{+\infty}\int_{-\infty}^{x}(x-y)\,dF(x)\,dF(y)-\int_{-\infty}^{+\infty}\int_{x}^{+\infty}(y-x)\,dF(x)\,dF(y)\right].$$

Rezultă

$$\Gamma=\frac{1}{2M}\cdot\frac{1}{2}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}|x-y|\,dF(x)\,dF(y)=\frac{1}{2M}\cdot\frac{\delta}{2}=\frac{1}{2}\left(\frac{\delta}{2M}\right).$$

In concluzie, aria de concentraţie este jumătate din coeficientul de concentraţie al lui Gini.

Coeficientul de variabilitate al lui Pearson şi coeficientul de concentraţie al lui Gini au un neajuns, care rezultă din faptul că sînt influenţaţi prea mult de M, momentul de ordinul întîi, şi se pot folosi puţin în problemele teoretice mai înaintate.

Pentru îndepărtarea acestui neajuns, putem considera ca variabilă a distribuţiei pe $y=\frac{x-M}{\sigma}$, pe care o vom numi *variabilă normată* sau *redusă*.

Se vede uşor că orice distribuţie cu variabila normată are momentul de ordinul întîi egal cu zero şi dispersia egală cu unu

$$\int_{-\infty}^{+\infty}\left(\frac{x-M}{\sigma}\right)dF(x)=\frac{1}{\sigma}(M-M)=0,$$

$$\int_{-\infty}^{+\infty}\left(\frac{x-M}{\sigma}\right)^2 dF(x)=\frac{1}{\sigma^2}(M_2-M)^2=1.$$

Distribuţiile exprimate cu ajutorul variabilei normate pot fi mai uşor comparate din punct de vedere al formei, normalităţii, oblicităţii etc.

Capitolul IV

Distribuţii cu două dimensiuni

§ 1. Consideraţii generale

Vom studia, în acest capitol, *distribuţiile cu două dimensiuni* sau *de două variabile*, distribuţii ale căror frecvenţe se referă la două caracteristici.

Distribuţiile de două variabile pot fi discrete sau continue în raport cu fiecare dintre variabile sau faţă de amîndouă în acelaşi timp.

Pentru reprezentarea distribuţiilor de două variabile în sistemul axelor rectangulare putem considera dreptunghiuri sau pătrate egale, sau nu, între ele. In acest caz, frecvenţele corespunzătoare se referă la centrul acestor dreptunghiuri sau pătrate.

Construcţii geometrice intuitive sau stereograme pentru astfel de distribuţii, deşi mai puţin folosite, se imaginează foarte simplu în spaţiul cu trei dimensiuni.

Aşa, de pildă, poliedrul frecvenţelor unei distribuţii de două variabile se va obţine, ridicînd o perpendiculară în centrul fiecărui dreptunghi sau pătrat pe planul său, luînd pe la un segment egal sau proporţional cu frecvenţa corespunzătoare şi unind apoi vîrfurile acestor segmente pentru a forma un poliedru cu o suprafaţă superioară simplă.

Scanned with OKEN Scanner

Histograma distribuţiei este o stereogramă compusă din mai multe paralelipipede construite pe fiecare din dreptunghiuri sau pătrate şi a căror înălţime este proporţională sau egală cu frecvenţele corespunzătoare.

Reprezentarea geometrică cu ajutorul poligonului frecvenţelor şi histogramei ne sugerează ideea, ca şi la distribuţiile cu o dimensiune, de a considera distribuţiile cu două dimensiuni ca prime aproximaţii ale unor distribuţii teoretice, care pot fi ele însile discontinue sau continue.

TABELA IV.1.1

Distribuţia numărului de elevi din două clase a VII-a în raport cu notele obţinute la matematică şi fizică

Note obţinute la fizică	Note obţinute la matematică				
	5	4	3	2	Total
5	2	2	0	0	4
4	1	6	1	0	18
3	0	4	15	4	33
2	0	0	10	13	3
Total	3	12	26	17	58

Vom da, în cele ce urmează, unele exemple de distribuţii cu două dimensiuni (tabela IV.1,1 şi IV.1,2).

Orînduirea frecvenţelor în tabela IV.1,2 prezintă o particularitate destul de interesantă.

Să presupunem că pentru o pereche (fiică, tată) s-au găsit respectiv taliile (64,3 ; 63,7). Acest caz va fi trecut cu o unitate în linia 64,5 şi coloana 63,5. Dacă rezultatul măsurătorii ne-a dat numere întregi, de pildă (65, 64), atunci vom trece frecvenţa 0,25 în fiecare din următoarele poziţii: (64,5 ; 63,5), (64,5 ; 64,5), (65,5 ; 63,5), (65,5 ; 64,5), prin (i, j) notînd poziţia din linia i şi coloana j. Dacă numai una din talii se exprimă în numere întregi, de pildă (64 ; 63,7), vom trece frecvenţa 0,5 în fiecare din poziţiile (63,5 ; 63,5), (64,5 ; 63,5). In modul acesta se obţin frecvenţe care nu sînt numere întregi, însă tabela este construită cu mai multă exactitate.

§ 2. Funcţia de distribuţie

Tabelele statistice considerate pînă acum cuprind date experimentale, rezultate ale experienţei, înregistrate în general ca frecvenţe.

Statistica matematică, după cum am arătat la distribuţiile cu o dimensiune, nu studiază fenomenele de masă numai din punct de vedere descriptiv, ci caută să formuleze reguli generale, în care aceste

Scanned with OKEN Scanner

TABELA IV.1.2

Distribuția a 1 376 tați și fiice după talie (date culese de Pearson și Lee, reproduse după cartea lui R. A. Fischer, Statistic Methods for research works)

Talia fiicelor	Talia taților																		
	58,5	59,5	60,5	61,5	62,5	63,5	64,5	65,5	66,5	67,5	68,5	69,5	70,5	71,5	72,5	73,5	74,5	75,5	Total
52,5	—	—	—	—	0,25	0,25	—	—	—	—	—	—	—	—	—	—	—	—	0,5
53,5	—	—	—	—	0,25	0,25	—	—	—	—	—	—	—	—	—	—	—	—	0,5
54,5	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
55,5	—	—	—	—	—	—	—	1	—	—	—	—	—	—	—	—	—	—	1
56,5	0,25	0,25	—	0,25	1,25	0,5	—	1	0,5	0,5	—	—	—	—	—	—	—	—	4,5
57,5	0,25	0,25	0,5	1,5	4,5	1	1,5	1,5	2,5	—	0,5	0,5	—	—	—	—	—	—	14,5
58,5	0,25	0,75	0,5	0,75	0,75	1	1,75	1,25	5	2,75	0,5	0,25	—	—	—	—	—	—	15,5
59,5	0,5	1	2	—	6	4,75	5	6,25	11,75	3,5	3,5	2	1,75	0,5	—	—	—	—	48,5
60,5	0,75	0,75	—	2,5	8	6,25	12,5	18,25	20,5	11	9	4,75	2,5	1,25	1,25	—	—	—	99
61,5	—	0,5	1,75	2	9,75	11,5	13	23,75	23,75	20,25	16,5	10,25	4,25	3	1,25	—	—	—	141,5
62,5	—	1	2,25	2	4,5	12	22,75	26	33	28 25	24,75	14,25	13,75	4 75	0,75	0,5	—	—	190,5
63,5	—	—	0,25	2	6	8,25	11	27,25	35,75	37,25	31,5	26,25	16,25	7,75	1,5	0,75	0,25	—	212
64,5	—	—	0,25	2,5	1,75	3,25	9,25	23	18,75	28,5	33	34,25	24,5	11,75	5,5	1	0,25	1	198,5
65,5	—	—	—	0,5	1	0,5	11	12,25	9,25	19,75	30	26,5	22,25	15	4,75	3,75	2	1	159,5
66,5	—	—	—	0,5	0,5	1,5	3,25	7,25	8,75	16	26,25	26,75	20,5	18,5	7,75	4,25	0,25	0,5	142,5
67,5	—	—	—	—	—	—	1	5,75	7	4	14,25	13,25	12	11,25	4,5	3,75	0,75	—	77,5
68,5	—	—	—	—	0,25	0,25	0,25	0,25	1,5	3	5,5	4,25	5,75	5,25	3,75	2,5	1,5	2	36
69,5	—	—	—	—	0,25	0,25	0,25	0,25	0,25	0,25	1	2,5	6,5	2,25	2,75	2	1	—	19,5
70,5	—	—	—	—	—	—	—	—	—	—	1,75	0,25	4,5	0,75	1,25	0,75	0,25	—	9,5
71,5	—	—	—	—	—	—	—	—	—	—	0,5	—	0,5	0,5	1,5	0,75	0,25	—	4
72,5	—	—	—	—	—	—	—	—	—	—	1	—	—	—	—	—	—	—	1
Total	2	4,5	7,5	14,5	45	51,5	12,5	155	17 8	175	199,5	166	135	82,5	36,5	20	6,5	4,5	1 376

Scanned with OKEN Scanner

fenomene se încadrează. Din acest punct de vedere fiecare rezultat al unei experienţe efective trebuie considerat ca una dintre valorile posibile cu probabilitate dată, ale unei variabile aleatorii. De aceea în stereogramele sau tabelele unei repartiţii cu două variabile, frecvenţele pot fi empirice sau teoretice, deci probabilităţi propriu-zise. Să considerăm de pildă un vector $V(x, y)$ de componente x şi y. V este o variabilă aleatorie vectorială, adică există o probabilitate corespunzătoare evenimentului.

$$a_1 < x \leqq a_2, \quad b_1 < y \leqq b_2,$$

pe care-l notăm sub formă vectorială

$$a < V \leqq b,$$

unde a şi b reprezintă vectorii $a(a_1, a_2)$, $b(b_1, b_2)$.

Această probabilitate, care se va nota $P(a \leqq V \leqq b)$, se exprimă cu ajutorul funcţiei de distribuţie de două variabile $F(\xi, \eta)$ definită prin egalitatea

$$F(\xi, \eta) = P(x \leqq \xi, \ y \leqq \eta). \qquad \text{(IV.2,1)}$$

Proprietăţile generale ale acestei funcţii sînt arătate în cap. I. Menţionăm numai următoarele.

1) Funcţia de distribuţie este nedescrescătoare în raport cu fiecare din variabilele ξ, η.

1') Funcţia de distribuţie ia valori cuprinse între 0 şi 1

$$F(\xi, \eta) = 0, \quad \text{pentru} \quad \xi = -\infty \quad \text{şi} \quad \eta = -\infty,$$

$$F(\xi, \eta) = 1, \quad \text{pentru} \quad \xi = \infty \quad \text{şi} \quad \eta = \infty.$$

2) Este monotonă, nedescrescătoare

$$F(\xi', \eta') \geqq F(\xi, \eta)$$

dacă

$$\xi' \geqq \xi, \quad \eta' \geqq \eta.$$

§ 3. Distribuţii discrete

Vectorul aleator $V(x, y)$ poate lua un număr finit sau o infinitate numărabilă de valori

$$v_{jk} = v_{jk}(a_j, b_k) \quad (j = 1, 2, \ldots; \ k = 1, 2, \ldots)$$

cu probabilităţi date $p_{jk} \geq 0$, astfel încît să avem

$$\sum_j \sum_k p_{jk} = 1.$$

Dacă $V(x, y)$ ia un număr finit de valori, se poate folosi următoarea tabelă care pune în evidenţă valorile v_{jk} şi probabilităţile respective (tabela IV.3,1).

TABELA IV.3,1

x \ y	b_1	b_2	. . . .	b_n	
a_1	p_{11}	p_{12}	. . . .	p_{1n}	$p_{1.}$
a_2	p_{21}	p_{22}	. . . .	p_{2n}	$p_{2.}$
a_m	p_{m1}	p_{m2}	. . . .	p_{mn}	$p_{m.}$
	$p_{.1}$	$p_{.2}$	. . . .	$p_{.n}$	

In tabelă sînt m linii şi n coloane. In total sînt $m+n$ alinieri, înţelegînd prin cuvîntul aliniere o linie sau o coloană.
Am notat

$$\sum_{k=1}^{n} p_{jk} = p_{j.} \quad (j=1, 2, \ldots, m),$$

$$\sum_{j=1}^{m} p_{jk} = p_{.k} \quad (k=1, 2, \ldots, m). \qquad \text{(IV.3,1)}$$

Rezultă

$$\sum_{j=1}^{m} p_{j.} = 1, \qquad \sum_{j=1}^{n} p_{.j} = 1.$$

$p_{j.}$ şi $p_{.j}$ sînt *probabilităţi marginale* şi anume $p_{j.}$ reprezintă probabilitatea ca $x = a_j$, oricare ar fi y, iar $p_{.j}$ probabilitatea egalităţii

$y=b_j$, oricare ar fi x, prin x și y notînd componentele vectorului aleator $V(x, y)$.
Să notăm prin $P(x=a_i \mid y=b_j)$ *probabilitatea condiționată* a egalității $x=a_i$, știind că $y=b_j$.

Avem

$$P(x=a_i \mid y=b_j)=\frac{P(x=a_i,\ y=b_j)}{P(y=b_j)}=\frac{p_{ij}}{p_{.j}}. \qquad \text{(IV.3,2)}$$

In mod analog

$$P(y=b_i \mid x=a_j)=\frac{P(x=a_j,\ y=b_i)}{P(x=a_j)}=\frac{p_{ji}}{p_{j.}}. \qquad \text{(IV.3,3)}$$

Dacă evenimentele $x=a_i$ și $y=b_j$ sînt independente, avem pentru orice i și j

$$P(x=a_i,\ y=b_j)=P(x=a_i)\cdot P(y=b_j)=p_{i.}p_{.j}$$

sau

$$p_{ij}=p_{i.}p_{.j}. \qquad \text{(IV.3,4)}$$

Reciproc, dacă această relație are loc pentru oricare $i=1, 2, \ldots, m$ și $j=1, 2, \ldots, n$, rezultă că cele două componente x și y ale vectorului V sînt independente între ele.
Dăm cîteva exemple de distribuții discrete cu două variabile.

Exemplul 1. Intr-o urnă sînt 50 bile albe, 25 bile roșii și 25 bile negre, identice în afară de culoare. Se fac din urnă patru extracții succesive, înapoindu-se de fiecare dată bila extrasă în urnă. Notăm cu x numărul de bile albe ieșite din aceste patru extracții și cu y numărul de bile roșii ieșite din același număr de extracții. Este evident că x și y nu pot lua decît valorile 0, 1, 2, 3, 4. Probabilitatea de a obține $x=a$, $y=b$ este egală cu

$$\frac{4!}{a!\,b!\,(4-a-b)!}\cdot\frac{1}{2^{8-a}},$$

în conformitate cu formula (IV.2,2).
Rezultă tabela IV.3.2

Exemplul 2. Intr-o urnă sînt 6 bile albe și 9 bile negre. Se fac din urnă șase extracții succesive, fără a pune înapoi bila extrasă. Notăm cu x numărul de bile albe obținut în primele trei extracții și cu y numărul de bile albe obținut din ultimele trei extracții.
Avem tabela IV.3,3.
Probabilitățile p_{ij} sînt calculate cu ajutorul formulelor stabilite la distribuția hipergeometrică.

TABELA IV.3,2

Numărul de bile roşii ieşite în patru extracţii succesive

Numărul de bile albe ieşite în patru extracţii succesive x \ y	0	1	2	3	4	
0	ρ	4ρ	6ρ	4ρ	ρ	16ρ
1	8ρ	24ρ	24ρ	8ρ	0	64ρ
2	24ρ	48ρ	24ρ	0	0	96ρ
3	32ρ	32ρ	0	0	0	64ρ
4	16ρ	0	0	0	0	16ρ
	81ρ	108ρ	54ρ	12ρ	ρ	1

TABELA IV.3,3 $\rho = \frac{1}{256}$

x \ y	0	1	2	3	
0	0,016 79	0,075 52	0,075 52	0,016 79	0,184 62
1	0,075 52	0,226 58	0,151 04	0,021 57	0,474 71
2	0,075 52	0,151 04	0,064 73	0,005 39	0,296 68
3	0,016 79	0,021 57	0,005 39	0,000 20	0,043 95
	0,184 62	0,474 71	0,296 68	0,043 95	1

Este de notat simetria tabelei.

In cazul general, să admitem că în urnă sînt A bile albe şi N bile negre, că în primele $\alpha+\beta$ extracţii am obţinut $x=\alpha$ bile albe şi în următoarele $\alpha'+\beta'$ extracţii $y=\alpha'$ bile albe. Probabilitatea ca să avem $x=\alpha$, $y=\alpha'$ este

$$\frac{C_A^{\alpha}\cdot C_N^{\beta}}{C_{A+N}^{\alpha+\beta}}\cdot\frac{C_{A-\alpha}^{\alpha'}\cdot C_{N-\beta}^{\beta'}}{C_{A+N-\alpha-\beta}^{\alpha'+\beta'}}=\frac{A!\,N!\,(\alpha+\beta)!\,(\alpha'+\beta')!\,(A+N-\alpha-\beta-\alpha'-\beta')!}{\alpha!\,\beta!\,\alpha'!\,\beta'!\,(A-\alpha-\alpha')!\,(N-\beta-\beta')!\,(A+N)!}.$$

Probabilitatea ca să avem $x=\alpha'$, în primele $\alpha'+\beta'$ extracţii şi $y=\alpha$ în următoarele $\alpha+\beta$ extracţii este

$$\frac{C_A^{\alpha'}C_N^{\beta'}}{C_{A+N}^{\alpha'+\beta'}}\cdot\frac{C_{A-\alpha'}^{\alpha}\cdot C_{N-\beta}^{\beta'}}{C_{A+N-\alpha'-\beta'}^{\alpha+\beta}}=\frac{A!\,N!\,(\alpha+\beta)!\,(\alpha'+\beta')!\,(A+N-\alpha-\beta-\alpha'-\beta')!}{\alpha!\,\beta!\,\alpha'!\,\beta'!\,(A-\alpha-\alpha')!\,(N-\beta-\beta')!\,(A+N)!}.$$

După cum se vede, cele două probabilităţi sînt egale.

Exemplul 3. Să considerăm un lanţ Markov, simplu şi constant, putînd lua valorile $a_1, a_2, \ldots, a_m$ cu probabilităţile de trecere

$$P(x_n=a_j \mid x_{n-1}=a_i)=p_{ij} \qquad (i, j=1, 2, \ldots, m),$$

$$p_{ij}\geq 0, \qquad \sum_{j=1}^{m} p_{ij}=1.$$

Notăm cu x prima variabilă a lanţului şi cu y variabila de rangul n a lanţului. Notăm cu p_k probabilitatea egalităţii $x=a_k$ $(k=1, 2, \ldots, m)$.

Rezultă tabela IV.3,4

TABELA IV.3,4

	a_1	a_2	. . . . .	a_m	
a_1	$p_1P_{11}^{(n)}$	$p_1P_{12}^{(n)}$	. . . . .	$p_1P_{1m}^{(n)}$	P_1
a_2	$p_2P_{21}^{(n)}$	$p_2P_{22}^{(n)}$	. . . . .	p_2P_{2m}	P_2
⋮	⋮	⋮	⋯	⋮	⋮
a_m	$p_mP_{m1}^{(n)}$	$p_mP_{m2}^{(n)}$	. . . . .	$p_mP_{mm}^{(n)}$	P_m
	$P_1^{(n)}$	$P_2^{(n)}$	. . . . .	$P_m^{(n)}$	

Prin

$$P_k^{(n)}=\sum_{j=1}^{m} p_jP_{jk}^{(n)},$$

am notat probabilitatea a priori a egalităţii $y=a_k$, oricare ar fi valorile luate de primele $(n-1)$ variabile ale lanţului.

Introducem notaţia

$$M_{rs}=M(x^r y^s). \qquad \text{(IV.3,5)}$$

M_{rs} este, prin urmare, valoarea medie a produsului $x^r y^s$

$$M_{rs}=\sum_{j=1}^{m}\sum_{k=1}^{n} p_{jk}a_j^r b_k^s. \qquad \text{(IV.3,6)}$$

Scanned with OKEN Scanner

Păstrînd aceeaşi nomenclatură ca pentru distribuţiile cu o singură variabilă, vom putea spune că M_{rs} este un *moment de ordinul* $r+s$ al distribuţiei noastre sau momentul de ordin (r, s).

Avem evident $M_{00}=1$.

Momentele de ordinul întîi sînt

$$M_{10}=\sum_{j=1}^{m}\sum_{k=1}^{n}p_{jk}a_j=\sum_{j=1}^{m}a_j\sum_{k=1}^{n}p_{jk}=\sum_{j=1}^{m}p_{j.}\,a_j,$$
$$M_{01}=\sum_{j=1}^{m}\sum_{k=1}^{n}p_{jk}b_k=\sum_{k=1}^{n}b_k\sum_{j=1}^{m}p_{jk}=\sum_{k=1}^{m}p_{.k}b_k. \qquad \text{(IV.3,7)}$$

De aici înainte vom nota, pentru simplificare,

$$M_{10}=m_1,\quad M_{01}=m_2.$$

Punctul din plan avînd coordonatele (m_1, m_2) poartă numele de *centru de greutate* al distribuţiei. El joacă acelaşi rol ca şi media aritmetică pentru distribuţiile cu o dimensiune.

Momentele de ordinul al doilea sînt

$$M_{20}=\sum_{j=1}^{m}\sum_{k=1}^{n}p_{jk}a_j^2=\sum_{j=1}^{m}p_{j.}\,a_j^2,$$
$$M_{11}=\sum_{j=1}^{m}\sum_{k=1}^{n}p_{jk}a_jb_k, \qquad \text{(IV.3,8)}$$
$$M_{02}=\sum_{j=1}^{m}\sum_{k=1}^{n}p_{jk}b_k^2=\sum_{k=1}^{n}p_{.k}b_k^2.$$

In cazul particular, cînd x şi y sînt independente între ele, sînt îndeplinite relaţiile (IV.3,4) şi rezultă

$$M_{11}=\sum_{j=1}^{m}\sum_{k=1}^{n}p_{jk}a_jb_k=\sum_{j=1}^{m}\sum_{k=1}^{n}p_{j.}p_{.k}a_jb_k=\sum_{j=1}^{m}p_{j.}\,a_j\sum_{k=1}^{n}p_{.k}b_k=m_1\cdot m_2.$$

Momentele centrate sînt definite prin egalitatea

$$m_{rs}=M[(x-m_1)^r(y-m_2)^s,$$
$$m_{rs}=\sum_{j=1}^{m}\sum_{k=1}^{n}p_{jk}(a_j-m_1)^r\cdot(b_k-m_2)^s.$$

Scanned with OKEN Scanner

Rezultă imediat pentru primele momente

$$m_{10}=\sum_{j=1}^{m}\sum_{k=1}^{n}p_{jk}\,(a_j-m_1)=0,$$

$$m_{01}=\sum_{j=1}^{m}\sum_{k=1}^{n}p_{jk}\,(b_k-m_2)=0,$$

$$m_{20}=\sum_{j=1}^{m}\sum_{k=1}^{n}p_{jk}\,(a_j-m_1)^2=\sum_{j=1}^{m}p_{j\cdot}(a_j-m_1)^2,$$

$$m_{11}=\sum_{j=1}^{m}\sum_{k=1}^{n}p_{jk}\,(a_j-m_1)\,(b_k-m_2),$$

$$m_{02}=\sum_{j=1}^{m}\sum_{k=1}^{n}p_{jk}\,(b_k-m_2)^2=\sum_{k=1}^{n}p_{\cdot k}\,(b_k-m_2)^2.$$

$m_{20}=\sigma_1^2$ reprezintă dispersia lui x, iar $m_{02}=\sigma_2^2$ dispersia lui y. m_{11} este *covarianța* componentelor x și y.
În cazul cînd x și y sînt independente, $m_{11}=0$.
Funcția caracteristică $\varphi(t,\ u)$ este definită prin relația

$$\varphi(t,\ u)=M(e^{(tx+uy)\,i}),$$

unde t și u sînt variabile reale

$$\varphi(t,\ u)=\sum_{j=1}^{m}\sum_{k=1}^{n}p_{jk}e^{(a_jt+b_ku)\,i}. \qquad \text{(IV.3,9)}$$

Derivatele parțiale ale lui $\varphi(t,\ u)$ în raport cu t și u în punctul $t=0$, $u=0$ ne dau momentele M_{rs}, în conformitate cu formulele (II.4,4).

§ 4. Valori medii condiționate

Notăm prin $M(x\,|\,y=b_j)$ valoarea medie a lui x, condiționată de egalitatea $y=b_j$ și prin $M(y\,|\,x=a_j)$ valoarea medie a lui y, condiționată de $x=a_j$.
Ținînd seama de relațiile (IV.3,2), și (IV.3,3) rezultă imediat

$$M(x\,|\,y=b_j)=\sum_{i=1}^{m}a_i\,P(x=a_i\,|\,y=b_j)=\frac{\sum_{i=1}^{m}p_{ij}\,a_i}{p_{\cdot j}},$$

$$M(y\,|\,x=a_j)=\sum_{j=1}^{n}b_j\,P(y=b_j\,|\,x=a_i)=\frac{\sum_{j=1}^{n}p_{ij}\,b_j}{p_{i\cdot}} \qquad \text{(IV.4.1)}$$

După cum se vede, $M(x|y=b_j)$ este o funcție de valorile b_j. Pentru fiecare valoare a lui b_j există o valoare corespunzătoare $a_j=M(x|y=b_j)$, $(j=1, 2, \ldots, n)$. Această funcție poartă numele de *regresiunea mediei lui x asupra lui y.*

Punctele de coordonate (a_j, b_j) $(j=1, 2, \ldots, n)$ dau, prin unirea lor prin segmente de linii drepte, *linia de regresiune a lui x asupra lui y.*

In mod analog $M(y|x=a_i)$ este o funcție de a_i. Pentru fiecare valoare a lui a_i există o valoare corespunzătoare $\beta_i=M(y|x=a_i)$ $(i=1, 2, \ldots, m)$, care ne dă regresiunea mediei lui y asupra lui x. In loc să considerăm media condiționată a valorilor lui y, putem să folosim și alte medii. Să notăm astfel prin $\xi(x|y=b_j)$ mediana valorilor lui x, condiționată de egalitatea $y=b_j$. Mărimea $\xi(x|y=b_j)$ este o funcție de b_j care poartă numele de *regresiunea medianei lui x asupra lui y.*

In mod similar, $\xi(y|x=a_i)$ reprezintă mediana valorilor lui y condiționată de egalitatea $x=a_i$. Privită ca funcție de a_i, ea poartă numele de regresiune a medianei lui y asupra lui x.

Regresiunea mediei și regresiunea medianei se bucură de anumite proprietăți extremale specifice mediei și medianei.

Să considerăm astfel expresia

$$M_2(x-\alpha|y=b_i),$$

care reprezintă momentul de ordinul al doilea al lui $x-\alpha$, condiționat de $y=b_j$. Știm că acest moment ia cea mai mică valoare pentru

$$\alpha_j=M(x|y=b_j).$$

Prin urmare, dacă ne punem problema determinării unei funcții $\alpha=\varphi(b)$, astfel ca expresia

$$M_2(x-\varphi(b_j)|y=b_j)$$

să devină minimă, $\varphi(b)$ este regresiunea mediei lui x asupra lui y. In mod analog, momentul condiționat de ordinul al doilea

$$M_2(y-\psi(a_i)|x=a_i)$$

Scanned with OKEN Scanner

devine minim pentru

$$\psi(a_i)=M(y\mid x=a_i).$$

Consideraţii analoge putem face şi pentru regresiunea medianei. Am văzut că mediana face minimă suma abaterilor absolute ale unei distribuţii. Prin urmare, dacă vom considera expresia

$$M(x-\varphi(b_j)\mid y=b_j),$$

minimul ei este atins pentru

$$\varphi(b_j)=\xi(x\mid y=b_j).$$

In mod analog, funcţia

$$M(y-\psi(a_i)\mid x=a_i)$$

devine minimă pentru

$$\psi(a_i)=\xi(y\mid x=a_i).$$

§ 5. Curbe de regresiune parabolică

Să considerăm o distribuţie discontinuă, de două variabile, ale cărei variabile x şi y pot lua, respectiv, valorile $a_i\,(i=1, 2, \ldots, m)$ şi $b_j\,(j=1, 2, \ldots, n)$ cu probabilităţile $P(x=a_i,\ y=b_j)=p_{ij}$.

Ne propunem să determinăm parametrii arbitrari α şi β astfel ca expresia

$$M(y-\alpha x-\beta)^2 \qquad \text{(IV.5,1)}$$

să fie minimă.

In baza definiţiei sale, expresia $M(y-\alpha x-\beta)^2$ se mai poate scrie

$$\sum_{i=1}^{m}\sum_{j=1}^{n} p_{ij}\,(b_j-\alpha a_i-\beta)^2. \qquad \text{(IV.5,2)}$$

Atunci, problema se reduce la determinarea minimului expresiei (IV.5,2). Trebuie, pentru aceasta, să anulăm derivatele parţiale de ordinul întîi în raport cu α şi β ale acestei expresii, adică

$$\begin{aligned}\sum_{i=1}^{m}\sum_{j=1}^{n} p_{ij}\,a_i(b_j-\alpha a_i-\beta)&=0,\\ \sum_{i=1}^{m}\sum_{j=1}^{n} p_{ij}\,(b_j-\alpha a_i-\beta)&=0.\end{aligned} \qquad \text{(IV.5,3)}$$

Relațiile (IV.5,3) sînt echivalente cu

$$M_{11}-\alpha M_{20}-\beta m_1=0,$$
$$m_2-\alpha m_1-\beta=0,$$

ținînd seama de notațiile adoptate (cap. III, § 10), de unde rezultă sistemul de ecuații

$$\begin{aligned}\alpha M_{20}+\beta m_1&=M_{11},\\ \alpha m_1+\beta&=m_2.\end{aligned}\qquad\text{(IV.5,4)}$$

Rezolvînd acest sistem, obținem următoarele valori pentru α și β

$$\alpha=\frac{M_{11}-m_1m_2}{M_{20}-m_1^2},\quad \beta=\frac{m_2M_{20}-m_1M_{11}}{M_{20}-m_1^2}.$$

Folosind relațiile cunoscute

$$\sigma_1^2=M_{20}-m_1^2,\quad \sigma_2^2=M_{02}-m_2^2,$$
$$\rho=\frac{M_{11}-m_{11}}{\sigma_1\sigma_2},$$

soluțiile de mai sus devin

$$\alpha=\frac{\rho\sigma_2}{\sigma_1},\quad \beta=\frac{m_2\sigma_1-m_1\rho\sigma_2}{\sigma_1}.$$

Pentru aceste valori ale lui α și β, expresia $M(y-\alpha x-\beta)^2$ este în adevăr minimă, deoarece se știe că dispersia unei variabile aleatorii în raport cu o valoare arbitrară a variabilei este minimă, dacă se ia ca valoare arbitrară media variabilei.

In cazul de față, dispersia variabilei $y-\alpha x-\beta$ în raport cu originea este minimă, dacă media $\sum_{i=1}^{m}\sum_{j=1}^{n}p_{ij}(b_j-\alpha a_i-\beta)=0$, condiție care este, de altfel, îndeplinită de ultima dintre egalitățile (IV.5,3).

Am obținut, astfel, dreapta de regresiune

$$y-\frac{\rho\sigma_2}{\sigma_1}x-\frac{m_2\sigma_1-m_1\rho\sigma_2}{\sigma_1}=0$$

sau

$$y-m_2=\frac{\rho\sigma_2}{\sigma_1}(x-m_1).\qquad\text{(IV.5,5)}$$

Scanned with OKEN Scanner

In mod analog, dacă ne-am fi propus să determinăm parametrii arbitrari α şi β în aşa fel ca expresia

$$M(x-\alpha y-\beta)^2$$

să fie minimă, am fi obţinut dreapta de regresiune

$$y-m_2=\frac{\rho\sigma_1}{\sigma_2}(x-m_1), \qquad \text{(IV.5,6)}$$

Cele două drepte, definite prin ecuaţiile (IV.5,5) şi (IV.5,6), sînt *dreptele de regresiune clasice* care, după cum se vede, trec prin centrul de greutate (m_1, m_2) al distribuţiei.

In general, aceste drepte de regresiune sînt diferite, afară de cazul cînd

$$\frac{\rho\sigma_1}{\sigma_2}=\frac{\rho\sigma_2}{\sigma_1}$$

sau

$$\sigma_1=\sigma_2.$$

Prin urmare, cele două drepte de regresiune vor coincide, dacă dispersiile σ_1 şi σ_2 sînt egale.

Cunoscînd valorile parametrilor α şi β pentru care expresia (IV.5,1) este minimă, vom putea să calculăm acest minim.

Astfel, minimul expresiei (IV.5,1) va fi

$$M\left[y-m_2-\frac{\rho\sigma_2}{\sigma_1}(x-m_1)\right]^2=M(y-m_2)^2-2\frac{\rho\sigma_2}{\sigma_1}M(x-m_1)(y-m_2)+$$

$$+\frac{\rho^2\sigma_2^2}{\sigma_1^2}M(x-m_1)^2=\sigma_2^2-\frac{2\rho\sigma_2}{\sigma_1}\rho\sigma_1\sigma_2+\frac{\rho^2\sigma_2^2}{\sigma_1^2}\sigma_1^2=\sigma_2^2(1-\rho^2)\ (\rho\leqslant 1).$$

In acelaşi fel, minimul expresiei $M(x-\alpha y-\beta)^2$ va fi egal cu

$$M\left[y-m_2-\frac{\rho\sigma_1}{\sigma_2}(x-m_1)\right]^2=\sigma_1^2(1-\rho^2)\ (\rho\leqslant 1).$$

Dăm o *interpretare geometrică a acestei probleme de minim*. In acest scop, să reprezentăm grafic (fig. IV.5,1) cele două drepte de regresiune, ale căror ecuaţii sînt scrise mai sus.

Minimul expresiei $M(y-\alpha x-\beta)^2$ adică $\sigma_2^2(1-\rho^2)$ reprezintă media pătratelor segmentelor MM', care unesc punctele $M(a_i, b_j)$ ale

distribuției cu intersecțiile M' ale ordonatelor punctelor M cu dreapta de regresiune (IV.5,5).

Cu alte cuvinte, pe dreapta de regresiune a cărei ecuație este

$$y-m_2=\frac{\rho\sigma_2}{\sigma_1}(x-m_1),$$

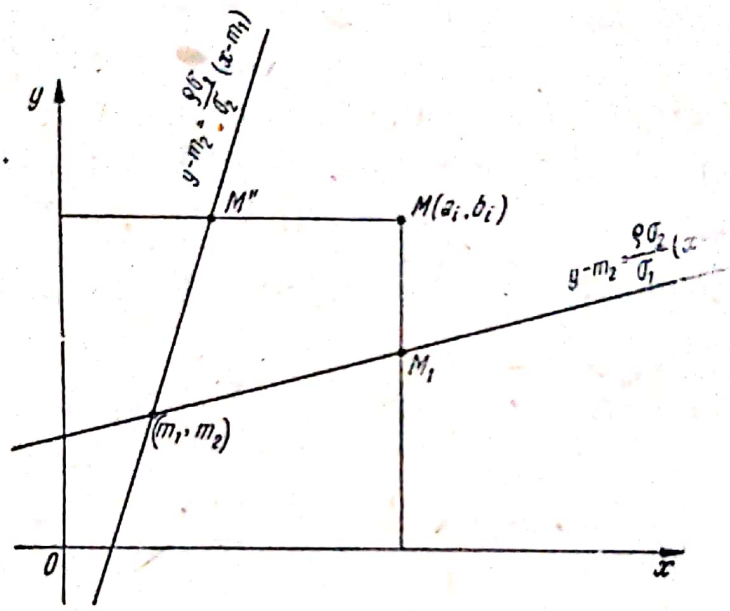


Fig. IV.5,1

sînt așezate toate punctele M' pentru care suma pătratelor distanțelor lor la punctele M ale distribuției, distanțe măsurate pe ordonatele punctelor M, este minimă.

De asemenea $\sigma_1^2(1-\rho^2)$, minimul expresiei $M(x-\alpha y-\beta)^2$ va reprezenta media pătratelor segmentelor MM'', care unesc punctele $M(a_i, b_j)$ ale distribuției cu intersecțiile M'' ale paralelelor duse prin punctele M la axa x-lor cu dreapta de regresiune (IV.5,6).

Dreapta de regresiune a cărei ecuație este

$$y-m_2=\frac{\rho\sigma_1}{\sigma_2}(x-m_2)$$

conține, deci, toate punctele M'' astfel că suma pătratelor distanțelor lor la punctele M ale distribuției, distanțe măsurate pe paralele duse din punctele M la axa x-lor, este minimă.

Deci, dreptele de regresiune se apropie cel mai mult de distribuția dată, în sensul pe care l-am arătat.

In cazul cînd $\sigma_1^2(1-\rho^2)=\sigma_2^2(1-\rho^2)=0$, ceea ce are loc numai dacă coeficientul de corelaţie $\rho=\pm 1$, atunci toate punctele distribuţiei se găsesc respectiv, pe fiecare din dreptele de regresiune clasică.
Coeficientul ρ, întîlnit mai sus, este prin definiţie, dat de expresia

$$\rho=\frac{M(x-m_1)(y-m_2)}{\sqrt{M(x-m_1)^2}\sqrt{M(y-m_2)^2}}.$$

Acest coeficient este cunoscut în statistică sub numele de *coeficient de corelaţie* şi se bucură de proprietăţile de mai jos.
1) Coeficientul de corelaţie ρ satisface inegalităţile

$$-1\leqslant\rho\leqslant 1.$$

In adevăr, să considerăm următoarea formă pătratică definită pozitivă

$$\sum_{i=1}^{m}\sum_{j=1}^{n}p_{ij}[\lambda(a_i-m_1)+\mu(b_j-m_2)]^2=\lambda^2\sum_{i=1}^{m}\sum_{j=1}^{n}p_{ij}(a_i-m_1)^2+$$

$$+2\lambda\mu\sum_{i=1}^{m}\sum_{j=1}^{n}p_{ij}(a_i-m_1)(b_j-m_2)+\mu^2\sum_{i=1}^{m}\sum_{j=1}^{n}p_{ij}(b_j-m_2)^2=$$

$$=\sigma_1^2\lambda^2+2\lambda\mu\sigma_1\sigma_2+\sigma_2^2\mu^2\geqslant 0.$$

Forma pătratică considerată, fiind definită pozitivă, discriminantul ei este pozitiv.
Prin urmare,

$$\begin{vmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{vmatrix}=\sigma_1^2\sigma_2^2(1-\rho^2)\geqslant 0.$$

De unde

$$\rho^2\leqslant 1,$$

sau

$$-1\leqslant\rho\leqslant 1.$$

2) Dacă coeficientul de corelaţie $\rho=\pm 1$, atunci între variabilele x şi y ale distribuţiei există o dependenţă liniară.
In acest caz discriminantul formei pătratice, definită pozitivă, de mai sus este nul. In acest caz, după cum se ştie, este necesar şi suficient ca parametrii λ şi μ să verifice relaţia liniară.

$$\lambda(a_i-m_1)+\mu(b_j-m_2)=0 \qquad (i,\ j=1,2,\ldots,m).$$

Scanned with OKEN Scanner

Valorile a_i și b_j luate de variabilele x și y ale distribuției verifică, deci, o relație liniară și deci între cele două variabile există dependența liniară

$$\lambda(x-m_1)+\mu(y-m_2)=0.$$

Pe dreapta definită de această ecuație se vor găsi toate punctele $P(a_i, b_j)$ ale distribuției; distribuția bidimensională se reduce la o distribuție unidimensională.

Parametrii λ și μ din această ecuație se pot determina ușor.

Discriminantul formei pătratice definită pozitivă fiind nul întrucît $\rho=\pm 1$, avem

$$\frac{\lambda}{\mu}=\frac{-\rho\sigma_1\sigma_2\pm\sqrt{\sigma_1^2\sigma_2^2(\rho^2-1)}}{\sigma_1^2}=\mp\frac{\sigma_2}{\sigma_1}.$$

In mod analog,

$$\frac{\mu}{\lambda}=\frac{-\rho\sigma_1\sigma_2\pm\sqrt{\sigma_1^2\sigma_2^2(\rho^2-1)}}{\sigma_2^2}=\mp\frac{\sigma_1}{\sigma_2}.$$

Obținem, astfel, următoarele drepte pe care se află punctele distribuției

$$y-m_2=\frac{\sigma_2}{\sigma_1}(x-m_1) \qquad (\rho=1),$$

$$y-m_2=-\frac{\sigma_2}{\sigma_1}(x-m_1) \qquad (\rho=-1),$$

respectiv

$$x-m_1=\frac{\sigma_1}{\sigma_2}(y-m_2) \qquad (\rho=1),$$

$$x-m_1=-\frac{\sigma_1}{\sigma_2}(y-m_2) \qquad (\rho=-1).$$

Acestea sînt dreptele de regresiune (IV.5,5) și (IV.5,6) găsite mai înainte, în cazul particular, cînd $\rho=\pm 1$.

Dacă ρ este foarte apropiat de ± 1, vom putea presupune distribuția examinată cu toate punctele apropiate de dreapta de regresiune.

3) Dacă variabilele x și y ale distribuției sînt independente între ele, atunci coeficientul de corelație este nul $\rho=0$.

Scanned with OKEN Scanner

Proprietatea rezultă imediat, deoarece în acest caz

$$M[(x-m_1)(y-m_2)]=M(x-m_1)\,M(y-m_2)=0.$$

Dacă, însă $\rho=0$, nu putem trage concluzia că există o dependență slabă între variabilele x și y ale distribuției. Dimpotrivă, exemplul următor ne dovedește contrariul. Să considerăm tabela de corelație

y \ x	-1	0	1	
0	0	$\frac{1}{2}$	0	$\frac{1}{2}$
1	$\frac{1}{4}$	0	$\frac{1}{4}$	$\frac{1}{2}$
	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	1

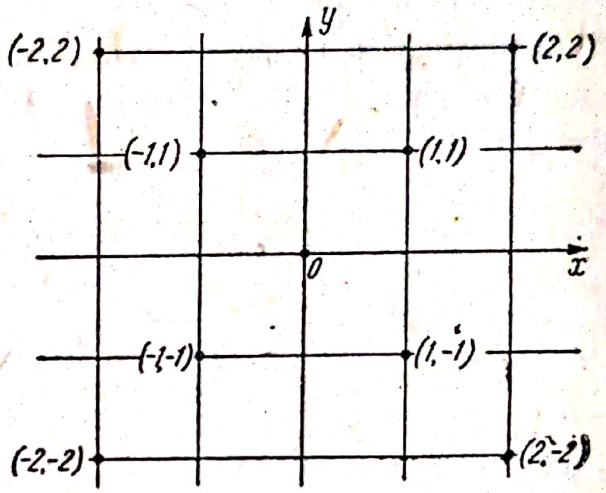


Fig. IV.5,2

Să observăm că punctele distribuției sînt așezate pe parabola $y=x^2$. De aici, deducem

$$m_1=(-1)\cdot\frac{1}{4}+0\cdot\frac{1}{2}+1\cdot\frac{1}{4}=0,$$

$$m_2=0\cdot\frac{1}{2}+1\cdot\frac{1}{2}=\frac{1}{2},$$

$$M(x-m_1)(y-m_2)=M(xy)-m_1m_2=(-1)\cdot\frac{1}{4}+1\cdot\frac{1}{4}=0$$

și, prin urmare, $\rho=0$.

În timp ce coeficientul de corelație este nul, între variabilele x și y dependența este directă, funcțională, $y=x^2$.

Cu atît mai mult dacă $\rho=0$ nu rezultă că variabilele x și y sînt independente între ele, cum se poate constata din exemplul de mai jos. Fie o distribuție formată din opt puncte ca în fig. IV.5,2.

În fiecare din punctele distribuției, masa p_{ij} $(i,\ j=1,2,\ldots,8)$ este egală, evident, cu $\frac{1}{8}$.

Se constată ușor că

$$m_1 = m_2 = 0,$$

$$M(x - m_1)(y - m_2) = 0$$

și deci

$$\rho = 0.$$

Calculînd masa din punctul (2,2), de pildă, găsim

$$p_{22} = \frac{1}{8}.$$

In același timp, avem

$$p_{2\cdot} = \frac{1}{4}, \quad p_{\cdot 2} = \frac{1}{4},$$

adică

$$p_{22} \neq p_{2\cdot} p_{\cdot 2}.$$

Așadar, dacă coeficientul de corelație $\rho = 0$, nu reiese în mod necesar independența variabilelor x și y.

De aceea, cînd coeficientul de corelație este nul se spune că variabilele x și y ale distribuției sînt *necorelate*, ceea ce, în lumina exemplului precedent cînd între y și x este dependența funcțională $y = x^2$, nu poate reprezenta o calificare foarte precisă.

Coeficientul de corelație ρ indică deci numai într-o măsură ce nu poate fi considerată satisfăcătoare legătura dintre variabilele x și y. Numai în cazul cînd el ia valorile limită ± 1 putem vorbi despre o situație categorică: există o dependență liniară între cele două variabile x și y.

Intr-un mod mai general, ne-am putea propune problema următoare.

Fiind dată o distribuție discontinuă ale cărei variabile x și y pot lua respectiv valorile $a_i\ (i, 1, 2, \ldots, m)$ și $b_j\ (j = 1, 2, \ldots, n)$ cu probabilitățile p_{ij}, să determinăm parametrii arbitrari $\alpha_h\ (h = 0, 1, \ldots, r)$ astfel ca expresia

$$M(y - \alpha_0 x^r - \alpha_1 x^{r-1} - \ldots - \alpha_r)^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} p_{ij} (b_j - \alpha_0 a_i^r - \alpha_1 a_i^{r-1} - \ldots - \alpha_r)^2 \tag{IV.5,7}$$

să fie minimă.

Anulînd derivatele parţiale de ordinul întîi în raport cu $\alpha_0, \alpha_1, \ldots, \alpha_r$ ale expresiei (IV.5,7), aşa cum am procedat şi mai înainte, obţinem sistemul de ecuaţii

$$\begin{aligned}
&M_{r,0}\alpha_0 + M_{r-1,0}\alpha_1 + \ldots + \alpha_r = m_2,\\
&M_{r+1,0}\alpha_0 + M_{r,0}\alpha_1 + \ldots + m_1\alpha_r = M_{11},\\
&M_{r+2,0}\alpha_0 + M_{r+1,0}\alpha_1 + \ldots + M_{2,0}\alpha_r = M_{21},\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&M_{2r,0}\alpha_0 + M_{2r-1,0}\alpha_1 + \ldots + M_{r,0}\alpha_r = M_{r1}.
\end{aligned} \tag{IV.5,8}$$

Acest sistem de $r+1$ ecuaţii cu $(r+1)$ necunoscute $(\alpha_0, \alpha_1, \ldots, \alpha_r)$ se poate rezolva întotdeauna, deoarece determinantul

$$\Delta = \begin{vmatrix} M_{r,0} & M_{r-1,0} \ldots & 1\\ M_{r+1,0} & M_{r,0} \ldots & m_1\\ M_{r+2,0} & M_{r+1,0} \ldots & M_{2,0}\\ \ldots & \ldots & \ldots\\ M_{2r,0} & M_{2r+1,0} \ldots & M_{r,0} \end{vmatrix} > 0, \tag{IV.5,9}$$

ceea ce s-a arătat într-una din teoremele anterioare (cap. II, § 2). Vom avea $\Delta = 0$, numai dacă numărul m al punctelor distribuţiei date este mai mic decît r şi în acest caz sistemul nu are soluţii.

Rezultă, deci că $\alpha_0, \alpha_1, \ldots, \alpha_r$ sînt unic determinate şi astfel putem găsi curba de regresiune, care va fi o parabolă de gradul r.

Valorile lui $\alpha_0, \alpha_1, \ldots, \alpha_r$ obţinute din sistemul (IV.5,8) corespund valorii minime a expresiei (IV.5,7) deoarece, după cum am arătat şi în problema de mai înainte, această expresie are efectiv un minim. Problema de faţă cuprinde evident, ca un caz particular problema de minim precedentă.

In adevăr, pentru $r=1$, sistemul (IV.5,8) devine sistemul (IV.5,4), iar curba de regresiune, care în cazul general este o parabolă de gradul r, devine o dreaptă a cărei ecuaţie este (IV.5,5).

Să presupunem acum că regresiunea medie a lui y asupra lui x este polinomul de gradul r

$$\varphi(x) = \beta_0 x^r + \beta_1 x^{r-1} + \ldots + \beta_r.$$

Ne interesează modul cum se leagă coeficienții $\beta_k (k=0, 1, \ldots, r)$ de valorile medii ale distribuției considerate.

Potrivit definiției sale, regresiunea medie a lui y asupra lui x se scrie

$$M(y \mid x=a_i)=\varphi(a_i)$$

sau

$$\frac{\sum_{j=1}^{n} p_{ij} b_j}{\sum_{j=1}^{n} p_{ij}} = \beta_0 a_i^r + \beta_1 a_i^{r-1} + \ldots + \beta_r,$$

de unde, eliminînd numitorul și înmulțind cu a_i^p, obținem

$$\sum_{j=1}^{n} p_{ij} a_i^p b_j = \beta_0 \sum_{j=1}^{n} p_{ij} a_i^{p+r} + \beta_1 \sum_{j=1}^{n} p_{ij} a_i^{p+r-1} + \ldots + \beta_r \sum_{j=1}^{n} a_i^p.$$

In sfîrșit, însumînd această egalitate și în raport cu indicele i de la 0 la m deducem

$$M_{p,1}=\beta_0 M_{p+r,0} + \beta_1 M_{p+r-1,0} + \ldots + \beta_r M_{p,0} \quad (p=0, 1, \ldots). \quad \text{(IV.5,10)}$$

Pentru valorile lui $p=0, 1, \ldots, r$, vom regăsi sistemul de ecuații (IV.5,8) și, deci, urmează că

$$\beta_0=\alpha_0, \quad \beta_1=\alpha_1, \quad \beta_2=\alpha_2, \ldots, \beta_r=\alpha_r.$$

De aici, rezultă proprietatea următoare:

Dacă regresiunea mediei lui y asupra lui x este un polinom de gradul r, atunci acest polinom coincide cu polinomul care reprezintă regresiunea parabolică de același grad.

In particular, dacă regresiunea mediei lui y asupra lui x este un polinom de gradul întîi, atunci acest polinom va coincide cu ecuația dreptei de regresiune clasică.

Problema de minim examinată mai sus nu este unică. Ne putem pune, de exemplu, după o idee a lui Gini, problema următoare:

Fiind dată o distribuție discretă de două variabile x și y, care pot lua respectiv valorile $a_i (i=1, 2, \ldots, m)$ și $b_j (j=1, 2, \ldots, n)$ cu probabilitățile p_{ij}, să se determine în planul distribuției o dreaptă D, astfel ca expresia $M(d^2)$ să fie minimă. Prin d înțelegem distanța de la un punct al distribuției date la dreapta D.

Fie $M(a_i, b_j)$ un punct al distribuţiei şi

$$x\cos\alpha+y\sin\alpha-p=0 \qquad \text{(IV.5,11)}$$

ecuaţia dreptei D.

Atunci, distanţa de la punctul $M(a_i, b_j)$ la dreapta D va fi

$$a_i\cos\alpha+b_j\sin\alpha-p=0$$

şi, deci, vom avea

$$M(d^2)=\sum_{i=1}^{m}\sum_{j=1}^{n}p_{ij}(a_i\cos\alpha+b_j\sin\alpha-p)^2.$$

Astfel fiind, urmează să determinăm parametrii α şi β în aşa fel ca expresia

$$\sum_{i=1}^{m}\sum_{j=1}^{n}p_{ij}(a_i\cos\alpha+b_j\sin\alpha-p)^2 \qquad \text{(IV.5,12)}$$

să fie minimă.

Anulînd derivata acesteia în raport cu p, obţinem

$$\sum_{i=1}^{m}\sum_{j=1}^{n}p_{ij}(a_i\cos\alpha+b_j\sin\alpha-p)=0$$

sau

$$m_1\cos\alpha+m_2\sin\alpha=p.$$

Pentru această valoare a lui p, expresia (IV.5,12) este minimă, deoarece derivata ei de ordinul al doilea are semnul pozitiv

$$2\sum_{i=1}^{m}\sum_{j=1}^{n}p_{ij}=2>0.$$

Ţinînd seamă de valoarea lui p, ecuaţia (IV.5,11) a dreptei D devine

$$(x-m_1)\cos\alpha+(y-m_2)\sin\alpha=0. \qquad \text{(IV.5,13)}$$

După cum se vede, această ecuaţie trece prin centrul de greutate (m_1, m_2) al distribuţiei date.

Ecuaţia (IV.5,13) înlocuieşte ecuaţia (IV.5,11) a dreptei D şi, prin urmare, vom scrie

$$M(d^2)=\sum_{i=1}^{m}\sum_{j=1}^{n}p_{ij}[(a_i-m_1)\cos\alpha+(b_j-m_2)\sin\alpha]^2.$$

Scanned with OKEN Scanner

Problema pusă se reduce, astfel, la determinarea parametrului α, pentru care expresia

$$\sum_{i=1}^{m}\sum_{j=1}^{n} p_{ij}\left[(a_i-m_1)\cos\alpha+(b_j-m_2)\sin\alpha\right]^2=$$

$$=\sigma_1^2\cos^2\alpha+2\rho\sigma_1\sigma_2\sin\alpha\cos\alpha+\sigma_2^2\sin^2\alpha \qquad \text{(IV.5,14)}$$

este minimă.

In acest scop, să considerăm în planul distribuției elipsa definită prin ecuația

$$\sigma_1^2X+2\rho\sigma_1\sigma_2XY+\sigma_2^2Y^2=1, \qquad \text{(IV.5,15)}$$

care, în urma transformării de variabile în coordonate polare

$$X=r\cos\alpha,$$

$$Y=r\sin\alpha,$$

devine

$$\sigma_1^2\cos^2\alpha+2\rho\sigma_1\sigma_2\sin\alpha\cos\alpha+\sigma_2^2\sin^2\alpha=\frac{1}{r^2},$$

adică expresia (IV.5,14) de mai sus.

Observăm că această expresie ia o valoare minimă cînd r este cel mai mare, adică cînd r coincide cu semiaxa mare a elipsei.

Aşadar, pentru aflarea minimului expresiei (IV.5,14) va trebui să calculăm semiaxa mare a elipsei a cărei ecuație este (IV.5,15). Calcule elementare ne arată că semiaxa mare a acestei elipse este egală cu

$$\sqrt{\frac{2}{\sigma_1^2+\sigma_2^2-\sqrt{(\sigma_1^2-\sigma_2^2)^2+4\rho^2\sigma_1^2\sigma_2^2}}}$$

şi, deci, minimul expresiei $M(d^2)$ este

$$\frac{1}{2}\left[\sigma_1^2+\sigma_2^2-\sqrt{(\sigma_1^2-\sigma_2^2)^2+4\rho^2\sigma_1^2\sigma_2^2}\right]. \qquad \text{(IV.5,16)}$$

Se poate constata uşor că acest minim este mai mic decît minimele precedente

$$\sigma_1^2(1-\rho^2) \quad \text{şi} \quad \sigma_2^2(1-\rho^2).$$

Va trebui, pentru aceasta, să arătăm, de pildă, că

$$\sigma_1^2(1-\rho^2)\geq\frac{1}{2}\left[\sigma_1^2+\sigma_2^2-\sqrt{(\sigma_1^2-\sigma_2^2)^2+4\rho^2\sigma_1^2\sigma_2^2}\right].$$

de unde deducem

$$\sqrt{(\sigma_1^2-\sigma_2^2)^2+4\rho^2\sigma_1^2\sigma_2^2} \geqq \sigma_1^2(2\rho^2-1)+\sigma_2^2$$

sau, ridicînd la pătrat şi efectuînd toate reducerile, găsim

$$\sigma_1^2(\rho^2-1) \leqq 0,$$

ceea ce este adevărat.

La fel, se arată că minimul (IV.5,16) este mai mic decît minimul $\sigma_2^2(1-\rho^2)$.

Determinînd pe α din ecuaţia

$$\sigma_1^2\cos^2\alpha+2\rho\sigma_1\sigma_2\sin\alpha\cos\alpha+\sigma_2^2\sin^2\alpha=$$

$$=\frac{1}{2}\left[\sigma_1^2+\sigma_2^2-\sqrt{(\sigma_1^2-\sigma_2^2)+4\rho^2\sigma_1^2\sigma_2^2}\right],$$

dreapta D este acum cunoscută.

Rezultă, deci, că distanţele d de la punctele $M(a_i, b_j)$ ale distribuţiei la dreapta D sînt, respectiv, mai mici decît distanţele MM' şi MM'', despre care am vorbit mai sus, în afară de cazul cînd $\rho^2=1$. In acest caz, distanţele d sînt egale cu zero şi pe dreapta D se găsesc toate punctele distribuţiei.

Minimul $\frac{1}{2}\left[\sigma_1^2+\sigma_2^2-\sqrt{(\sigma_1^2-\sigma_2^2)^2+4\rho^2\sigma_1^2\sigma_2^2}\right]$ există întotdeauna, cu excepţia cazului cînd elipsa se reduce la un cerc. Acesta este, evident, un caz particular care are loc, numai dacă dispersiile σ_1 şi σ_2 sînt egale, iar coeficientul de corelaţie $\rho=0$.

Toate aceste consideraţii le-am făcut în cazul distribuţiilor discrete, care se întîlnesc mai frecvent în aplicaţiile practice ale statisticii matematice. Ele rămîn, însă, valabile şi în cazul distribuţiilor continue, schimbînd sumele obişnuite prin integrale ordinare.

De exemplu, în cazul distribuţiei normale cu două dimensiuni, a cărei funcţie de frecvenţă este

$$f(x, y)=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\cdot e^{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-a_1)^2}{\sigma_1^2}-2\rho\frac{(x-a_1)(y-a_2)}{\sigma_1\sigma_2}+\frac{(y-a_2)^2}{\sigma_2^2}\right]},$$

vom avea

$$m_1 = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} x f(x, y)\, dx\, dy = a_1,$$

$$m_2 = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} y f(x, y)\, dx\, dy = a_2,$$

$$m_{20} = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} (x-a_1)^2 f(x, y)\, dx\, dy = \sigma_1^2,$$

$$m_{11} = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} (x-a_1)(y-a_2) f(x, y)\, dx\, dy = \rho\sigma_1\sigma_2,$$

$$m_{02} = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} (y-a_2)^2 f(x, y)\, dx\, dy = \sigma_2^2,$$

de unde rezultă că dreptele de regresiune ale acestei distribuții sînt

$$y - a_2 = \frac{\rho\sigma_2}{\sigma_1}(x - a_1)$$

și

$$y - a_2 = \frac{\rho\sigma_1}{\sigma_2}(x - a_1).$$

De asemenea, regresiunile mediilor condiționate sînt

$$M(x \mid y) = \frac{\int_{-\infty}^{+\infty} x f(x, y)\, dx}{\int_{-\infty}^{+\infty} f(x, y)\, dy} = a_2 + \frac{\rho\sigma_1}{\sigma_2}(x - a_1)$$

și

$$M(y \mid x) = \frac{\int_{-\infty}^{+\infty} y f(x, y)\, dy}{\int_{-\infty}^{+\infty} f(x, y)\, dx} = a_2 + \rho\frac{\sigma_2}{\sigma_1}(x - a_1),$$

adică tot niște drepte care, evident, după teoria generală, trebuie să coincidă cu dreptele de regresiune clasică.

Ca o observație generală, se constată că, pentru $\rho=0$, variabilele x și y sînt independente între ele, deoarece funcția de frecvență $f(x, y)$ de mai sus se poate pune sub forma

$$f(x, y)=\varphi(x)\cdot\psi(y).$$

§ 6. Coeficientul de contingență al lui Pearson

Fiind dată o distribuție discretă de două variabile x și y, care pot lua respectiv valorile $a_i (i=1, 2, \ldots, m)$ și $b_j (j=1, 2, \ldots, n)$ cu probabilitățile p_{ij}, să determinăm valoarea parametrului α, pentru care expresia

$$f^2=\frac{1}{\alpha}\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{(p_{ij}-p_{i\cdot}p_{\cdot j})^2}{p_{i\cdot}p_{\cdot j}} \qquad \text{(IV.6,1)}$$

să aibă cea mai mare valoare egală cu unu.
In acest caz, ne interesează maximul sumei

$$\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{(p_{ij}-p_{i\cdot}p_{\cdot j})^2}{p_{i\cdot}p_{\cdot j}}, \qquad \text{(IV.6,2)}$$

care se mai scrie, dezvoltînd,

$$\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{p_{ij}^2}{p_{i\cdot}p_{\cdot j}}-2\sum_{i=1}^{m}\sum_{j=1}^{n}p_{ij}+\sum_{i=1}^{m}\sum_{j=1}^{n}p_{i\cdot}p_{\cdot j}=\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{p_{ij}^2}{p_{i\cdot}p_{\cdot j}}-1,$$

avînd în vedere relațiile

$$\sum_{i=1}^{m}\sum_{j=1}^{n}p_{ij}=1,\quad \sum_{i=1}^{m}p_{i\cdot}=1 \quad \text{și} \quad \sum_{j=1}^{n}p_{\cdot j}=1.$$

Deoarece $\sum_{j=1}^{n}\frac{p_{ij}}{p_{i\cdot}}=1$, observăm că expresia

$$\sum_{j=1}^{n}\frac{p_{ij}^2}{p_{i\cdot}p_{\cdot j}}=\sum_{j=1}^{n}\frac{p_{ij}}{p_{i\cdot}}\cdot\frac{p_{ij}}{p_{\cdot j}}$$

poate fi interpretată ca valoarea medie a unei distribuții discrete a cărei variabilă ia valorile $\frac{p_{ij}}{p_{\cdot j}}$ $(j=1, 2, \ldots, n)$ cu probabilitățile $\frac{p_{ij}}{p_{i\cdot}}\geq 0$, indicele i fiind constant.

Insă, după cum am arătat (cap. II, § 2), valoarea medie este cuprinsă între cea mai mică şi cea mai mare valoare a variabilei şi, deci, avem inegalitatea

$$\sum_{j=1}^{n} \frac{p_{ij}}{p_{i.}} \cdot \frac{p_{.j}}{p_{.j}} \leq \frac{p_{ij'}}{p_{.j'}} = \frac{p_{ij'}}{\sum_{i=1}^{m} p_{ij'}} < 1,$$

unde am notat cu $\frac{p_{ij'}}{p_{.j'}}$ cea mai mare valoare a variabilei considerate.
Insumînd şi în raport cu indicele i, deducem

$$\sum_{i=1}^{m} \sum_{j=1}^{n} \frac{p_{ij}^2}{p_{i.}p_{.j}} < \sum_{i=1}^{m} 1 = m.$$

De unde rezultă că maximul expresiei (IV.6,2) este

$$m-1,$$

dacă $m<n$.
Maximul expresiei (IV.6,1), în aceste condiţii, va fi evident

$$\frac{1}{\alpha}(m-1).$$

In mod analog, dacă $m>n$, acest maxim va fi egal cu

$$\frac{1}{\alpha}(n-1).$$

Cum, din condiţiile problemei, maximul expresiei (IV.6,1) trebuie să fie egal cu unu, urmează că parametrul α va avea valoarea

$$m-1$$

respectiv

$$n-1.$$

Din motive de simetrie, s-a luat pentru parametrul α, valoarea

$$\sqrt{(m-1)(n-1)}.$$

Atunci expresia (IV.6,1) devine

$$f^2 = \frac{1}{\sqrt{(m-1)(n-1)}} \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{(p_{ij}-p_{i.}p_{.j})^2}{p_{i.}p_{.j}}, \qquad \text{(IV.6,3)}$$

expresie care poartă numele de *coeficient de contingenţă al lui Pearson.*

In cele ce urmează, vom arăta cîteva proprietăţi ale acestui coeficient.

1) Dacă variabilele x şi y sînt independente între ele, coeficientul de contingenţă al lui Pearson este egal cu zero, deoarece în baza relaţiei

$$p_{ij} = p_{i.}p_{.j},$$

rezultă

$$f^2 = 0.$$

2) Invers, dacă coeficientul de contingenţă $f^2 = 0$, variabilele x şi y sînt independente între ele.

In adevăr, dacă $f^2 = 0$, avem din (IV.6,3)

$$p_{ij} = p_{i.}p_{.j}.$$

3) Coeficientul de contingenţă al lui Pearson satisface inegalităţile

$$0 \leq f^2 = \frac{1}{\sqrt{(m-1)(n-1)}} \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{(p_{ij} - p_{i.}p_{.j})}{p_{i.}p_{.j}} < 1.$$

Astfel, dacă $m < n$, cea mai mare valoare a lui f^2 este

$$\frac{1}{\sqrt{(m-1)(n-1)}}(m-1) = \sqrt{\frac{m-1}{n-1}} < 1,$$

iar dacă $n < m$, această valoare este egală cu

$$\frac{1}{\sqrt{(m-1)(n-1)}}(n-1) = \sqrt{\frac{n-1}{m-1}} < 1.$$

Coeficientul de contingenţă este, evident, pozitiv, fiind o sumă de numere pozitive şi este egal cu zero cînd variabilele x şi y sînt independente între ele.

4) Dacă între variabilele x şi y există o dependenţă funcţională $y = f(x)$, atunci coeficientul de contingenţă al lui Pearson are valoarea 1.

O dependenţă funcţională $y = f(x)$ între variabilele x şi y înseamnă că fiecărei valori a_i $(i = 1, 2, \ldots, m)$ a variabilei x îi corespunde o valoare bine determinată b_j $(j = 1, 2, \ldots, n)$ a variabilei y.

Scanned with OKEN Scanner

Prin urmare, numărul relativ p_i al valorilor a_i este egal cu numărul relativ p_j al valorilor b_j. Mai mult chiar, numărul relativ p_{ij} al perechilor de valori (a_i, b_j) este egal cu fiecare dintre numerele relative $p_{i.}$ şi $p_{.j}$, adică

$$p_{i.}=p_{.j}=p_{ij}.$$

In acelaşi timp,

$$m=n.$$

Ca urmare a celor de mai sus, expresia (IV.6,3) a coeficientului de contingenţă al lui Pearson se reduce la

$$f^2=\frac{1}{m-1}\left(\sum_{i=1}^{m}\sum_{j=1}^{m}\frac{p_{ij}^2}{p_{i.}p_{.j}}-1\right)=\frac{1}{m-1}\left(\sum_{i=1}^{m}\sum_{j=1}^{m}\frac{p_{ij}}{p_{i.}}-1\right)=$$

$$=\frac{1}{m-1}\left(\sum_{i=1}^{m}1-1\right)=1.$$

Proprietatea reciprocă nu are loc.
Cu alte cuvinte, dacă coeficientul de contingenţă al lui Pearson are valoarea 1, nu se poate afirma că există o dependenţă funcţională $y=\varphi(x)$ între variabilele x şi y.

§ 7. Coeficientul maxim de corelaţie

După cum am arătat mai înainte, coeficientul de corelaţie ρ dă indicaţii în general slabe cu privire la dependenţa dintre variabilele x şi y, în afară de cazul cînd el ia una dintre valorile limită -1 sau $+1$. Atunci putem afirma că între variabilele x şi y există o corelaţie strînsă exprimată, aproape peste tot, ca o relaţie liniară. Din acest motiv, s-au determinat alte caracteristici care să dea indicaţii mai satisfăcătoare relativ la corelaţia dintre variabilele x şi y.
Dăm aici, ca un exemplu, definiţia şi proprietăţile *coeficientului maxim de corelaţie* al lui H. Gebelein, aducînd unele precizări necesare, care decurg în mod simplu din proprietăţile generale ale matricelor lanţurilor Markov, finite şi constante.
In acest scop, fie o distribuţie discretă de două variabile x şi y, care pot lua, respectiv, valorile a_i $(i=1, 2, \ldots, m)$ şi b_j $(j=1, 2, \ldots, n)$ şi $p_{ij}=\text{Prob}\,[x=a_i,\ y=b_j]$.

Avînd $f(x)$ și $g(x)$ două funcții regulate, să considerăm expresia

$$r^2=\frac{\left(\sum_{i=1}^{m}\sum_{j=1}^{n}p_{ij}f_ig_j\right)^2}{\left(\sum_{i=1}^{m}\sum_{j=1}^{n}p_{ij}f_i^2\right)\left(\sum_{i=1}^{m}\sum_{j=1}^{n}p_{ij}g_j^2\right)}, \qquad \text{(IV.7,1)}$$

unde

$$f_i=f(a_i) \quad \text{și} \quad g_j=g(b_j).$$

Ne propunem să determinăm funcțiile f și g astfel ca această expresie să fie maximă.

In baza inegalității lui Schwarz

$$\left[\sum_{i=1}^{m}\sum_{j=1}^{n}(\sqrt{p_{ij}}f_i)(\sqrt{p_{ij}}g_i)^2\right]^2\leq\sum_{i=1}^{m}\sum_{j=1}^{n}p_{ij}f_i^2\sum_{i=1}^{m}\sum_{j=1}^{n}p_{ij}g_j^2,$$

rezultă că întotdeauna avem

$$r^2\leq 1.$$

Fără a restrînge generalitatea problemei, putem să ne alegem funcțiile f și g astfel ca valorile lor medii să fie nule

$$\sum_{i=1}^{m}\sum_{j=1}^{n}p_{ij}f_i=0, \quad \sum_{i=1}^{m}\sum_{j=1}^{n}p_{ij}g_i=0$$

sau

$$\sum_{i=1}^{m}p_{i.}f_i=0, \quad \sum_{j=1}^{n}p_{.j}g_j=0. \qquad \text{(IV.7,2)}$$

Altfel spus, egalitățile (IV.7,2) sînt echivalente cu o schimbare a axelor de coordonate, luînd originea în punctul de coordonate respectiv egale cu valorile medii ale valorilor f_i și g_j.

Totodată, putem norma funcțiile f și g, ceea ce revine a scrie egalitățile

$$\sum_{i=1}^{m}\sum_{j=1}^{n}p_{ij}f_i^2=1, \quad \sum_{i=1}^{m}\sum_{j=1}^{n}p_{ij}g_j^2=1$$

sau

$$\sum_{i=1}^{m}p_{i.}f_i^2=1, \quad \sum_{j=1}^{n}p_{.j}g_j^2=1. \qquad \text{(IV.7,3)}$$

Scanned with OKEN Scanner

Atunci, expresia (IV.7,1), care definește coeficientul r, se reduce la forma mai simplă

$$r^2=\left(\sum_{i=1}^{m}\sum_{j=1}^{n}p_{ij}f_ig_j\right)^2, \qquad \text{(IV.7,4)}$$

Mai întîi să presupunem că funcția $g(y)$ este cunoscută. In acest caz, trebuie să alegem funcția $f(x)$ în așa fel, încît coeficientul r, dat acum de expresia (IV.7,4), să fie maxim.

Aplicînd din nou inegalitatea lui Schwarz, avem

$$\left[\sum_{i=1}^{m}\left(\sqrt{p_{i.}}f_i\right)\left(\frac{1}{\sqrt{p_{i.}}}\sum_{j=1}^{n}p_{ij}g_j\right)\right]^2\leqslant\sum_{i=1}^{m}p_{i.}f_i^2\sum_{i=1}^{m}\frac{1}{p_{i.}}\left(\sum_{j=1}^{n}p_{ij}g_j\right)^2=$$

$$=\sum_{i=1}^{m}\frac{1}{p_{i.}}\left(\sum_{j=1}^{n}p_{ij}g_j\right)^2,$$

deoarece $\sum_{i=1}^{m}p_{i.}f_i=1$.

Prin urmare, dacă funcția $g(y)$ este cunoscută, maximul lui r^2 este

$$\sum_{i=1}^{m}\frac{1}{p_{i.}}\left(\sum_{j=1}^{n}p_{ij}g_j\right)^2.$$

Acest maxim, după cum se vede, va fi atins dacă

$$\frac{\sqrt{p_{i.}}f_i}{\frac{1}{\sqrt{p_{i.}}}\sum_{j=1}^{n}p_{ij}g_j}=\frac{1}{\lambda},$$

λ fiind o constantă, care se poate determina ușor.

Din relația de mai sus, deducem

$$\frac{1}{\lambda^2}=\frac{p_{i.}f_i^2}{\frac{1}{p_{i.}}\left(\sum_{j=1}^{n}p_{ij}g_j\right)^2}=\frac{\sum_{i=1}^{m}p_{i.}f_i^2}{\sum_{i=1}^{m}\frac{1}{p_{i.}}\left(\sum_{j=1}^{n}p_{ij}g_j\right)^2}=\frac{1}{\sum_{i=1}^{m}\frac{1}{p_{i.}}\left(\sum p_{ij}g_j\right)^2},$$

adică λ^2 este tocmai maximul expresiei (IV.7,4).

De aici, rezultă

$$f_i = \frac{1}{\lambda p_{i\cdot}} \sum_{j=1}^{n} p_{ij} g_j \qquad (i=1, 2, \ldots, m). \qquad \text{(IV.7,5)}$$

Am determinat, astfel, funcţia $f(x)$ în punctele $a_1, a_2, \ldots, a_m$ în funcţie de valorile funcţiei $g(y)$ în punctele $b_1, b_2, \ldots, b_j$.
Invers, dacă funcţia $f(x)$ este dată, atunci, procedînd ca mai sus şi notînd cu μ^2 maximul lui r^2, vom obţine

$$g_j = \frac{1}{\mu p_{\cdot j}} \sum_{i=1}^{m} p_{ij} f_i \qquad (j=1, 2, \ldots, n). \qquad \text{(IV.7,6)}$$

In modul acesta, se determină funcţia $g(y)$ în punctele $b_1, b_2, \ldots, b_j$ (în funcţie de valorile funcţiei $f(x)$ în punctele $a_1, a_2, \ldots, a_m$).
Dacă r^2 are un maxim, atunci el trebuie să fie egal cu

$$\lambda^2 = \mu^2.$$

In adevăr, fie M^2 acest maxim. Am stabilit mai înainte că $r^2 \leqq \lambda^2$ dacă funcţia $g(y)$ este dată. Alegînd funcţia $f(x)$ astfel ca $r^2 = M^2$, rezultă că $M^2 \leqq \lambda^2$.
Cum, însă, λ^2 este cea mai mare valoare a lui r^2, înseamnă că

$$M^2 = \lambda^2.$$

In mod analog, se poate arăta că şi

$$M^2 = \mu^2.$$

Aşadar, avem

$$\lambda^2 = \mu^2.$$

Din egalităţile (IV.7,5) şi (IV.7,6), deducem următorul sistem liniar omogen de $m+n$ ecuaţii cu $m+n$ necunoscute ($f_1, f_2, \ldots, f_m$; $g_1, g_2, \ldots, g_n$)

$$\lambda p_{i\cdot} f_i = \sum_{j=1}^{n} p_{ij} g_j \qquad (i=1, 2, \ldots, m),$$

$$\lambda p_{\cdot j} g_j = \sum_{i=1}^{m} p_{ij} f_i \qquad (j=1, 2, \ldots, n). \qquad \text{(IV.7,7)}$$

Scanned with OKEN Scanner

Pentru ca acest sistem liniar omogen să admită soluţii diferite de zero, trebuie

$$\begin{vmatrix} -\lambda p_1 & 0 & \dots & 0 & p_{11} & p_{12} & \dots & p_{1n} \\ 0 & -\lambda p_2 & \dots & 0 & p_{21} & p_{22} & \dots & p_{2n} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & -\lambda p_m & p_{m1} & p_{m2} & \dots & p_{mn} \\ p_{11} & p_{21} & \dots & p_{m1} & -\lambda p_{.1} & 0 & \dots & 0 \\ p_{12} & p_{22} & \dots & p_{m2} & 0 & -\lambda p_{.2} & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ p_{1n} & p_{2n} & \dots & p_{mn} & 0 & 0 & \dots & -\lambda p_{.n} \end{vmatrix} = 0.$$

Introducînd notaţiile

$$\frac{p_{ij}}{p_{i.}} = s_{ij}, \quad \frac{p_{ij}}{p_{.j}} = t_{ji},$$

ecuaţia de mai sus se va scrie

$$\begin{vmatrix} -\lambda & 0 \dots 0 & s_{11} & s_{12} \dots s_{1n} \\ 0 & -\lambda \dots 0 & s_{21} & s_{22} \dots s_{2n} \\ \dots & \dots & \dots & \dots \\ 0 & 0 \dots -\lambda & s_{m1} & s_{m2} \dots s_{mn} \\ t_{11} & t_{12} \dots t_{1m} & -\lambda & 0 \dots 0 \\ t_{21} & t_{22} \dots t_{2m} & 0 & -\lambda \dots 0 \\ \dots & \dots & \dots & \dots \\ t_{n1} & t_{n2} \dots t_{nm} & 0 & 0 \dots -\lambda \end{vmatrix} = 0. \qquad \text{(IV.7,8)}$$

Avem, evident, relaţiile

$$\sum_{j=1}^{n} s_{ij} = \sum_{j=1}^{n} \frac{p_{ij}}{p_i} = 1,$$

$$\sum_{i=1}^{m} t_{ij} = \sum_{i=1}^{m} \frac{p_{ij}}{p_{.j}} = 1.$$

Scanned with OKEN Scanner

Atunci, tabela de mai jos

$$\left\|\begin{array}{cccccc} 0 & 0 \ldots 0 & s_{11} & s_{12} \ldots s_{1n} \\ 0 & 0 \ldots 0 & s_{21} & s_{22} \ldots s_{2n} \\ \cdot & \cdot \quad \cdot \quad \cdot & \cdot & \cdot \quad \cdot \quad \cdot \\ 0 & 0 \ldots 0 & s_{m1} & s_{m2} \ldots s_{mn} \\ t_{11} & t_{12} \ldots t_{1m} & 0 & 0 \ldots 0 \\ t_{21} & t_{22} \ldots t_{2m} & 0 & 0 \ldots 0 \\ \cdot & \cdot \quad \cdot \quad \cdot & \cdot & \cdot \quad \cdot \quad \cdot \\ t_{n1} & t_{n2} \ldots t_{nm} & 0 & 0 \ldots 0 \end{array}\right\|$$

este o matrice stochastică finită, care se întîlneşte în studiul lanţurilor Markov finite şi constante.

In acest caz, toate rădăcinile ecuaţiei (IV.7,8) sînt în modul mai mici sau cel mult egale cu unu, adică satisfac relaţia

$$|\lambda| \leqq 1.$$

Dacă $m \geqq n$, se constată, în cazul normal, că această ecuaţie are $m-n$ rădăcini egale cu zero şi deci

$$m+n-(m-n)=2n$$

rădăcini diferite de zero.

Cu alte cuvinte, dacă p este cel mai mic dintre numerele m şi n, ecuaţia (IV.7,8) va avea $2p$ rădăcini diferite de zero. Mai mult, se poate arăta că toate aceste rădăcini sînt reale.

Presupunînd, astfel, că λ este o rădăcină complexă, atunci evident şi conjugata ei $\bar{\lambda}$ va fi o rădăcină a ecuaţiei (IV.7,8), iar o dată cu sistemul (IV.7,7) este veriticat şi sistemul liniar şi omogen

$$\begin{aligned} \bar{\lambda} p_{i.} \bar{f}_i &= \sum_{j=1}^{n} p_{ij} \bar{g}_j \qquad (i=1, 2, \ldots, m), \\ \bar{\lambda} p_{.j} \bar{g}_j &= \sum_{i=1}^{m} p_{ij} \bar{f}_i \qquad (j=1, 2, \ldots, n), \end{aligned} \tag{IV.7,9}$$

unde $\bar{f}_i$ şi $\bar{g}_j$ sînt, respectiv, conjugatele lui f_i şi g_j.

Inmulțind primele m ecuații ale sistemului (IV.7,7) cu $\bar{f}_i$, iar ultimele sale n ecuații cu $\bar{g}_j$, obținem, prin adunare,

$$\lambda\left(\sum_{i=1}^{m} p_{i.} f_i \bar{f}_i + \sum_{j=1}^{n} p_{.j} g_j \bar{g}_j\right) = \sum_{i=1}^{m}\sum_{j=1}^{n} p_{ij}(g_j \bar{f}_i + f_i \bar{g}_j). \quad \text{(IV.7,10)}$$

De asemenea, înmulțind aceleași ecuații ale sistemului (IV.7,9) respectiv cu f_i și g_j și adunînd, găsim

$$\bar{\lambda}\left(\sum_{i=1}^{m} p_{i.} f_i \bar{f}_i + \sum_{j=1}^{n} p_{.j} g_j \bar{g}_j\right) = \sum_{i=1}^{m}\sum_{j=1}^{n} p_{ij}(f_i \bar{g}_j + g_j \bar{f}_i). \quad \text{(IV.7,11)}$$

Scăzînd membru cu membru (IV.7,10) și (IV.7,11), deducem

$$(\lambda - \bar{\lambda})\left(\sum_{i=1}^{m} p_{i.} f_i \bar{f}_i + \sum_{j=1}^{n} p_{.j} g_j \bar{g}_j\right) = 0.$$

Observăm că expresia

$$\sum_{i=1}^{m} p_{i.} f_i \bar{f}_i + \sum_{j=1}^{n} p_{.j} g_j \bar{g}_j$$

este întotdeauna pozitivă, în afară de cazul cînd

$$f_i = g_j = 0,$$

probabilitățile $p_{i.}$ și $p_{.j}$ fiind presupuse diferite de zero.
Insă atunci tabela de corelație încetează de a mai fi o tabelă stochastică cu m linii și n coloane și urmează deci

$$\lambda = \bar{\lambda}.$$

Această egalitate nu poate avea loc decît dacă rădăcinile λ sînt reale, ceea ce ne-am propus să arătăm.
In sfîrșit, se mai poate arăta că ecuația (IV.7,8), o dată cu λ are ca rădăcină și pe $(-\lambda)$.
In adevăr, sistemul liniar și omogen (IV.7,7) este echivalent cu sistemul

$$(-\lambda)\, p_{i.} f_i = \sum_{j=1}^{n} p_{ij}(-g_j) \qquad (i = 1, 2, \ldots, m)$$

$$(-\lambda)\, p_{.j}(-g_j) = \sum_{i=1}^{m} p_{ij} f_i \qquad (j = 1, 2, \ldots, m),$$

Scanned with OKEN Scanner

care prin transformarea $h_j = -g_j$ devine

$$(-\lambda)\, p_{i.}\, f_i = \sum_{j=1}^{n} p_{ij}\, h_j \qquad (i=1,\ 2,\ \ldots,\ m)$$

$$(-\lambda)\, p_{.j}\, h_j = \sum_{i=1}^{m} p_{ij}\, f_i \qquad (j=1,\ 2,\ \ldots,\ n). \tag{IV.7,12}$$

Notînd cu $D(\lambda)$ determinantul sistemului (IV.7,7), se vede că $D(-\lambda)$ este determinantul sistemului (IV.7,12). Pentru ca cele două sisteme liniare şi omogene să aibă soluţii diferite de zero, este necesar şi suficient ca

$$D(\lambda)=0, \quad D(-\lambda)=0,$$

adică ecuaţia (IV.7,8) are ca rădăcini pe λ şi $(-\lambda)$.

In concluzie, ecuaţia (IV.7,8) sau, cu notaţiile de mai sus,

$$D(\lambda)=0$$

admite $2p$ rădăcini reale diferite de zero (p fiind cel mai mic număr dintre numerele m şi n), cuprinse în intervalul $[-1,\ 1]$, două cîte două egale şi de semne contrarii.

In continuare, să arătăm că putem reduce gradul ecuaţiei de la $2p$ la p.

Din (IV.7,7) obţinem, eliminînd pe g_j

$$\lambda p_{i.}\, f_i = \sum_{j=1}^{n} p_{ij}\, \frac{1}{\lambda p_{.j}} \sum_{k=1}^{m} p_{kj}\, f_k$$

sau

$$\lambda^2 f_i = \sum_{k=1}^{m} q_{ik}\, f_k \qquad (i=1,\ 2,\ \ldots,\ m), \tag{IV.7,13}$$

dacă notăm

$$q_{ik} = \sum_{j=1}^{n} \frac{p_{ij}\, p_{kj}}{p_{i.}\, p_{.j}}. \tag{IV.7,14}$$

Noul sistem liniar şi omogen (IV.7,13) ne conduce, astfel, la următoarea ecuaţie de gradul p în raport cu λ^2, presupunînd că $m<n$

$$|q_{ik}-\lambda^2\delta_{ik}|=\begin{vmatrix} q_{11}-\lambda^2 & q_{12} & \dots & q_{1m} \\ q_{21} & q_{22}-\lambda^2 & \dots & q_{2m} \\ \dots & \dots & \dots & \dots \\ q_{m1} & q_{m2} & \dots & q_{mm}-\lambda^2 \end{vmatrix}=0, \qquad \text{(IV.7,15)}$$

unde $\delta_{ik}=1$, dacă $i=k$ şi $\delta_{ik}=0$ dacă $i\neq k$.
Deoarece avem

$$\sum_{k=1}^{m} q_{ik}=\sum_{k=1}^{m}\sum_{j=1}^{n}\frac{p_{ij}p_{kj}}{p_{i.}\,p_{.j}}=\sum_{j=1}^{n}\frac{p_{ij}}{p_{i.}}=1, \quad q_{ik}\geqq 0,$$

rezultă că

$$\|q_{ik}\|=\begin{Vmatrix} q_{11} & q_{12}\dots q_{1m} \\ q_{21} & q_{22}\dots q_{2m} \\ \dots & \dots \\ q_{m1} & q_{m2}\dots q_{mm} \end{Vmatrix}$$

este o matrice stochastică.
Se ştie atunci că rădăcinile ecuaţiei (IV.7,15) verifică inegalitatea

$$\lambda^2\leqq 1.$$

Cum, însă, λ^2 este maximul lui r^2, definit de expresia (IV.7,4), înseamnă că acest maxim va fi dat de modulul celei mai mari dintre rădăcinile ecuaţiei

$$|q_{ik}-\lambda^2\delta_{ik}|=0,$$

în afară de rădăcina 1.
Coeficientul r astfel definit, prin cea mai mare dintre valorile lui λ^2 se numeşte *coeficient maxim de corelaţie.*
Între coeficientul maxim de corelaţie şi coeficientul de contingenţă al lui Pearson există o relaţie, care se exprimă prin următoarea

Teoremă. *Dacă*

$$1\geqq\lambda_1^2\geqq\lambda_2^2\geqq\dots\geqq\lambda_{m-1}^2$$

sînt rădăcinile ecuaţiei (IV.7,15)

$$|q_{ik}-\lambda^2\delta_{ik}|=0,$$

atunci, coeficientul maxim de corelație $r=\lambda_1$ *satisface inegalitatea*

$$\lambda_1^2 \geqq \sqrt{(m-1)(n-1)}\cdot f^2 \geqq f^2 \qquad (m \leqq n),$$

unde f este coeficientul de contingență al lui Pearson.

Dezvoltînd membrul întîi al ecuației (IV.7,15) după puterile descrescătoare ale lui λ^2, obținem

$$|q_{ik}-\lambda^2\delta_{ik}|=(-1)^m\lambda^{2m}+(-1)^{m-1}\sum_{i=1}^{m} q_{ii}\lambda^{m-2}+\ldots=0.$$

Pentru suma rădăcinilor, avem expresia

$$\lambda_1^2+\lambda_2^2+\ldots+\lambda_{m-1}^2=\sum_{i=1}^{m} q_{ii}-1=\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{P_{ij}^2}{p_{i\cdot}\,p_{\cdot j}}-1=$$

$$=\sqrt{(m-1)(n-1)}\cdot f^2, \qquad \text{(IV.7,16)}$$

ținînd seama de relația (IV.7,14) și de definiția coeficientului de contingență al lui Pearson.

Inlocuind în (IV.7,16) fiecare din rădăcinile $\lambda_2^2, \lambda_3^2, \ldots, \lambda_{m-1}^2$ cu cea mai mare rădăcină λ_1^2, rezultă

$$(m-1)\lambda_1^2 \geqq \sqrt{(m-1)(n-1)}\cdot f^2$$

sau

$$\lambda_1^2 \geqq \sqrt{\frac{n-1}{m-1}}\cdot f^2 \geqq f^2,$$

știind că $m \leqq n$.

Teorema este, deci, demonstrată.

Egalitatea dintre coeficientul maxim de corelație și coeficientul de contingență al lui Pearson va avea loc, dacă

$$m=n$$

și

$$\lambda_1^2=\lambda_2^2=\ldots=\lambda_{m-1}^2.$$

Aceasta înseamnă că ecuația

$$|q_{ik}-\lambda^2\delta_{ik}|=0$$

are o rădăcină multiplă de ordinul $m-1$, ceea ce după teoria lanțurilor Markov finite și constante, înseamnă că matricea este de $m-1$ ori decompozabilă și că, prin urmare, admite și pe 1 ca rădăcină multiplă de ordinul $m-1$. Cazul acesta este posibil, numai dacă

$$q_{ii}=1, \quad q_{ij}=0.$$

Incheiem acest paragraf, menționînd proprietățile deosebite ale coeficientului maxim de corelație, care rezultă ușor din expunerea de pînă aci.

1) Dacă variabilele x și y ale distribuției considerate sînt independente, coeficientul maxim de corelație este zero.

Aceasta rezultă din faptul că, în acest caz, coeficientul de contingență al lui Pearson este zero, și deci putem să scriem

$$\lambda_1^2+\lambda_2^2+\ldots+\lambda_{m-1}^2=\sqrt{(m-1)(n-1)}\cdot f^2=0,$$

ceeă ce este posibil, numai dacă

$$\lambda_1^2=\lambda_2^2=\ldots=\lambda_{m-1}^2=0.$$

2) Spre deosebire de cele stabilite în cazul coeficientului de corelație obișnuit, aici este adevărată și reciproca proprietății și anume, dacă coeficientul maxim de corelație este zero, variabilele x și y sînt independente între ele.

Dacă $\lambda_1^2=0$, avem evident și

$$\lambda_2^2=\lambda_3^2=\ldots=\lambda_{m-1}^2,$$

iar în baza egalității (IV.7,16) urmează imediat

$$f^2=0,$$

ceea ce înseamnă că variabilele x și y sînt independente între ele.

3) Dacă între variabilele x și y este o dependență funcțională, atunci $r^2=1$. Am arătat mai înainte că, dacă există o dependență funcțională între variabilele x și y, coeficientul de contingență al lui Pearson este 1, iar $m=n$. Proprietatea enunțată rezultă atunci imediat, deoarece din teorema de mai sus, avem

$$\lambda_1^2 \geqq f^2=1.$$

Intrucît cea mai mare rădăcină λ nu depășește valoarea unu, urmează că

$$r^2=\lambda_1^2=1.$$

Această proprietate are o consecinţă, care constă în aceea că şi celelalte rădăcini $\lambda_2^2, \lambda_3^2, \ldots, \lambda_{m-1}^2$, mai mici decît λ_1^2, sînt egale fiecare cu unu.
În adevăr, din relaţia (IV.7,16), obţinem

$$\lambda_1^2+\lambda_2^2+\ldots+\lambda_{m-1}^2=m-1,$$

deoarece $m=n$ şi $f^2=1$.
Cunoscînd, pe de o parte, că $\lambda_1^2=1$ este cea mai mare rădăcină, iar pe de altă parte, că toate rădăcinile sînt cuprinse între -1 şi $+1$, rezultă că ultima egalitate va avea loc numai dacă

$$\lambda_2^2=\lambda_3^2=\ldots=\lambda_{m-1}^2=1.$$

4) Dacă coeficientul maxim de corelaţie este 1, atunci există, în general, o dependenţă directă între x şi y, cu excepţia cazurilor cînd f_i şi g_j nu sînt diferite între ele.
Pornind de la egalitatea

$$f_ig_j=\frac{1}{2}\left[f_i^2+g_j^2-(f_i-g_j)^2\right],$$

deducem

$$r=\sum_{i=1}^{m}\sum_{j=1}^{n}p_{ij}f_ig_j=\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n}p_{ij}f_i^2+\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n}p_{ij}g_j^2-$$

$$-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n}p_{ij}(f_i-g_j)^2=1-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n}p_{ij}(f_i-g_j)^2\cdot$$

Dacă $r=1$, rezultă

$$\sum_{i=1}^{m}\sum_{j=1}^{n}p_{ij}(f_i-g)^2=0,$$

adică

$$f_i=g_j \qquad (p_{ij}\neq 0).$$

Prin urmare, există întotdeauna această dependenţă funcţională pentru tabela de corelaţie construită cu funcţiile f_i şi g_j. Mai mult chiar,

dacă valorile lui f_i și g_j sînt toate diferite între ele, atunci există o dependență funcțională și între valorile a_i și b_j deci, între valorile x și y.

§ 8. Calculul corelațiilor cu ajutorul dezvoltării după polinoame ortogonale

In studiul problemei momentelor, fiecărei distribuții îi corespunde un șir de polinoame ortogonale și normate. Aceste polinoame fiind strîns legate de distribuția dată, este natural să fie folosite pentru studiul ei. Fiecare dintre cele două distribuții marginale ale unei repartiții bidimensionale dă naștere la cîte o familie de polinoame ortogonale. Intocmai ca și pentru distribuțiile de o variabilă, este firesc să se folosească aceste polinoame la studiul diferitelor probleme în care intervin momentele distribuției și, deci, în particular și la studiul corelației.

Astfel fiind, ne propunem să calculăm coeficientul de contingență al lui Pearson, coeficientul de corelație și coeficientul maxim de corelație cu ajutorul dezvoltării după polinoamele ortogonale corespunzătoare.

Să considerăm, în acest sens, o tabelă de corelație ale cărei valori sînt a_i $(i=1,2,\ldots,m)$ și b_j $(j=1,2,\ldots,n)$ cu probabilitățile p_{ij}, precum și succesiunile de momente

$$S_0=1, \quad S_1=\sum_{i=1}^{m} p_{i.}\, a_i, \ldots, S_r=\sum_{i=1}^{m} p_{i.}\, a_i^r, \ldots,$$

$$T_0=1, \quad T_1=\sum_{j=1}^{n} p_{.j}\, b_j, \ldots, T_r=\sum_{j=1}^{n} p_{.j} b_j^r, \ldots,$$

corespunzătoare, respectiv, distribuțiilor marginale

$$p_{i.}a_i, \quad p_{.j}b_j.$$

Fiecare dintre aceste succesiuni de momente corespunde respectiv unui șir de polinoame ortogonale

$$\varphi_0(x), \quad \varphi_1(x), \ldots, \varphi_r(x), \ldots,$$

$$\psi_0(y), \quad \psi_1(y), \ldots, \psi_r(y), \ldots,$$

definite prin expresiile

$$\varphi_r(x)=\frac{(-1)^r}{\sqrt{D_r\cdot D_{r+1}}}\begin{vmatrix} 1 & x\dots x^r \\ 1 & S_1\dots S_{r-1} \\ S_1 & S_2\dots S_r \\ \cdot & \cdot\ \cdot\ \cdot\ \cdot \\ S_{r-1} & S_r\dots S_{2r-1} \end{vmatrix};$$

$$\psi_r(y)=\frac{(-1)^r}{\sqrt{\Delta_r\cdot\Delta_{r+1}}}\begin{vmatrix} 1 & y\dots y^r \\ 1 & T_1\dots T_{r-1} \\ T_1 & T_2\dots T_r \\ \cdot & \cdot\ \cdot\ \cdot\ \cdot \\ T_{r-1} & T_r\dots T_{2r-1} \end{vmatrix},$$

unde

$$D_r=\begin{vmatrix} 1 & S_1\dots S_{r-1} \\ S_1 & S_2\dots S_k \\ \cdot & \cdot\ \cdot\ \cdot\ \cdot \\ S_{r-1} & S_r\dots S_{2r-1} \end{vmatrix};$$

$$\Delta_r=\begin{vmatrix} 1 & T_1\dots T_{r-1} \\ T_1 & T_2\dots T_r \\ \cdot & \cdot\ \cdot\ \cdot\ \cdot \\ T_{r-1} & T_r\dots T_{2r-1} \end{vmatrix} \quad (r=1,2,\dots). \qquad \text{(IV.8,1)}$$

Ortogonalitatea este exprimată de relațiile

$$\sum_{i=1}^{m} p_{i.}\varphi_r(a_i)\varphi_s(a_i)=0;\ \sum_{j=1}^{n} p_{.j}\psi_r(b_j)\psi_s(b_j)=0 \qquad (r\neq s),$$
$$\sum_{i=1}^{m} p_{i.}\varphi_r^2(a_i)=1,\ \sum_{j=1}^{n} p_{.j}\psi_r(b_j)=1. \qquad \text{(IV.8.2)}$$

Să introducem în locul funcțiilor $f(x)$ și $g(y)$, considerate în paragraful precedent, polinoamele lui Lagrange $F(x)$ și $G(y)$, pentru care avem

$$F(a_i)=f(a_i);\quad G(b_j)=g(b_j) \qquad \begin{matrix}(i=1,2,\dots,m)\\(j=1,2,\dots,n).\end{matrix} \qquad \text{(IV.8,3)}$$

Polinoamele Lagrange $F(x)$ și $G(y)$ sînt de gradele $m-1$, respectiv $n-1$ și deci se pot exprima în funcție de primele m (respectiv n) polinoame ortogonale, adică

$$F(x)=\sum_{k=0}^{m-1} A_k \varphi_k(x); \quad G(y)=\sum_{k=0}^{n-1} B_k \psi_k(y), \qquad \text{(IV.8,4)}$$

A_k și B_k fiind niște coeficienți, care se pot determina precum urmează.

Pentru determinarea coeficienților A_k, de exemplu, înmulțim în prima dintre egalitățile (IV.8.4) cu $p_i \varphi_i(x)$.

Astfel, obținem

$$p_i \varphi_i(x) F(x)=\sum_{k=0}^{m-1} p_i A_k \varphi_k(x) \varphi_i(x).$$

nsumînd relativ la primul indice, avem

$$\sum_{x=a_1}^{a_m} p_i \varphi_i(x) F(x)=\sum_{k=0}^{m-1} A_k \sum_{x=a_1}^{a_m} p_i \varphi_k(x) \varphi_i(x).$$

In baza relațiilor de ortogonalitate (IV.8,2), termenii sumei duble din membrul al doilea al ultimei egalități sînt egali cu zero, pentru $k \neq i$. Rămîn, deci termenii acestei sume, pentru $k=i$, adică A_i.

Prin urmare, putem scrie

$$A_i=\sum_{x=a_1}^{a_m} p_i \varphi_i(x) F(x).$$

In mod analog, găsim

$$B_j=\sum_{y=b_1}^{b_n} p_j \psi_j(y) G(y).$$

O proprietate deosebită a acestor coeficienți constă în faptul că dau loc la egalitățile

$$\sum_{k=0}^{m-1} A_k^2=1; \quad \sum_{k=0}^{n-1} B_k^2=1.$$

Demonstrăm numai prima egalitate.

Ridicînd la pătrat în prima dintre egalitățile (IV.8,4), obținem

$$F^2(x)=\sum_{k=0}^{m-1} A_k^2 \varphi_k^2(x)+2\sum_{k \neq l} A_k A_l \varphi_k(x) \varphi_l(x).$$

Această egalitate este verificată pentru fiecare dintre valorile $a_1, a_2, \ldots, a_m$ ale variabilei x. O vom lua pentru o valoare oarecare a_i a variabilei x şi vom înmulţi în ea cu $p_{i.} \neq 0$. Insumînd apoi, după indicele i avem

$$\sum_{i=1}^{m} p_{i.} F^2(a_i) = \sum_{k=0}^{m-1} A_k^2 \sum_{i=1}^{m} p_{i.} \varphi_k^2(a_i) + 2 \sum_{k \neq l} A_k A_l \sum_{i=1}^{m} p_{i.} \varphi_k(a_i) \varphi_l(a_i).$$

Ţinînd seama de relaţiile de ortogonalitate (IV.8,2) şi de faptul că putem să alegem funcţia $F(x)$ astfel ca

$$\sum_{i=1}^{m} p_{i.} F^2(a_i) = 1,$$

rezultă

$$\sum_{k=0}^{m-1} A_k^2 = 1.$$

Să exprimăm, acum, probabilitatea p_{ij} cu ajutorul polinoamelor ortogonale $\varphi_r(x)$ şi $\psi_s(y)$. Ne propunem pentru aceasta să determinăm coeficienţii C_{rs} astfel ca

$$p_{ij} = \sum_{i=1}^{m} \sum_{j=1}^{n} p_{i.} p_{.j} C_{rs} \varphi_r(a_i) \psi_s(b_j). \qquad \text{(IV.8,5)}$$

Inmulţind în această egalitate cu $\varphi_t(a_i) \cdot \psi_u(b_j)$ şi însumînd după indicii i şi j de la 1 la m (respectiv de la 1 la n) vom avea

$$\sum_{i=1}^{m} \sum_{j=1}^{n} p_{ij} \varphi_t(a_i) \psi_u(b_j) =$$

$$= \sum_{r=0}^{m-1} \sum_{s=0}^{n-1} \sum_{i=1}^{m} \sum_{j=1}^{n} p_{i.} p_{.j} C_{rs} \varphi_r(a_i) \psi_s(b_j) \varphi_t(a_i) \psi_u(b_j) =$$

$$= \sum_r \sum_s C_{rs} \sum_i p_{i.} \varphi_r(a_i) \varphi_t(a_i) \sum_j p_{.j} \psi_s(b_j) \psi_u(b_j).$$

Se observă că toţi coeficienţii lui C_{rs} sînt nuli, în afară de cazul

$$r = t, \quad s = u.$$

De aici rezultă

$$C_{tu} = \sum_{i=1}^{m} \sum_{j=1}^{n} p_{ij} \varphi_t(a_i) \psi_u(b_j) \qquad \text{(IV.8,6)}$$

Scanned with OKEN Scanner

şi, în consecinţă, probabilitatea p_{ij} este determinată în funcţie de polinoamele $\varphi_r(x)$ şi $\psi_s(y)$.
In particular, din (IV.8,6) deducem

$$C_{00}=\sum_{i=1}^{m}\sum_{j=1}^{n}p_{ij}=1,$$

$$C_{0u}=C_{t0}=0.$$

O dată cu determinarea probabilităţii p_{ij}, în funcţie de polinoamele ortogonale, putem să trecem la rezolvarea efectivă a problemei ce ne-am propus.
Vom începe cu calcularea coeficientului de contingenţă al lui Pearson. Expresia (IV.8,5) se mai poate pune sub forma

$$\frac{p_{ij}}{p_{i.}p_{.j}}=\sum_{r=0}^{m-1}\sum_{s=0}^{n-1}C_{rs}\varphi_r(a_i)\psi_s(b_j).$$

Ridicînd la pătrat şi însumînd după indicii i şi j de la 1 la m (respectiv de la 1 la n) obţinem

$$\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{p_{ij}^2}{p_{i.}p_{.j}}=\sum_{i=1}^{m}\sum_{j=1}^{n}\left[\sum_{r=0}^{m-1}\sum_{s=0}^{n-1}C_{rs}\varphi_r(a_i)\psi_s(b_j)\right]^2 p_{i.}p_{.j}=$$

$$=\sum_{r=0}^{m-1}\sum_{s=0}^{n-1}C_{rs}^2\sum_{i=1}^{m}p_{i.}\varphi_r^2(a_i)\sum_{j=1}^{n}p_{.j}\psi_s^2(b_j)=\sum_{r=0}^{m-1}\sum_{s=0}^{n-1}C_{rs}^2=1+\sum_{r-1}^{m-1}\sum_{s=1}^{n-1}C_{rs}^2,$$

în conformitate cu relaţiile de ortogonalitate (IV.8,2).
Cum, însă, coeficientul de contingenţă al lui Pearson este dat de expresia

$$f^2=\frac{1}{\sqrt{(m-1)(n-1)}}\left(\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{p_{ij}^2}{p_{i.}p_{.j}}-1\right),$$

urmează că

$$f^2=\frac{\sum_{r=0}^{m-1}\sum_{s=0}^{n-1}C_{rs}^2}{\sqrt{(m-1)(n-1)}}.$$

Scanned with OKEN Scanner

Trecem la calculul coeficientului de corelaţie ρ, a cărui expresie, este

$$\rho=\frac{\sum_{i=1}^{m}\sum_{j=1}^{n}p_{ij}(a_i-m_1)(b_j-m_2)}{\sqrt{p_{i.}(a_i-m_1)^2}\cdot\sqrt{p_{.j}(b_j-m_2)^2}}.$$

Deducem din (IV.8,1)

$$\varphi_1(a_i)=-\frac{\begin{vmatrix}1 & a_i\\ 1 & S_1\end{vmatrix}}{\sqrt{\begin{vmatrix}1 & S_1\\ S_1 & S_2\end{vmatrix}}}=\frac{a_i-m_1}{\sigma_1};\quad \psi_1(b_j)=-\frac{\begin{vmatrix}1 & b_j\\ 1 & T_1\end{vmatrix}}{\sqrt{\begin{vmatrix}1 & T_1\\ T_1 & T_2\end{vmatrix}}}=\frac{b_j-m_2}{\sigma_2}$$

şi ρ devine

$$\rho=\frac{\sigma_1\sigma_2\sum_i\sum_j p_{ij}\varphi_1(a_i)\psi_1(b_j)}{\sigma_1\sqrt{p_{i.}\varphi_1(a_i)^2}\cdot\sigma_2\sqrt{p_{.j}\psi_1(b_j)^2}}=$$

$$=\sum_{r=0}^{m-1}\sum_{s=0}^{n-1}C_{rs}\sum_{i=1}^{m}p_{i.}\varphi_1(a_i)\varphi_r(a_i)\sum_{j=1}^{n}p_{.j}\psi_1(b_j)\psi_s(b_j),$$

deci

$$\rho=C_{11},$$

deoarece pentru $r\neq 1$ şi $s\neq 1$ toţi coeficienţii lui C_{rs} sînt nuli.
In sfirşit, să calculăm coeficientul maxim de corelaţie r.
Folosind relaţiile (IV.8,3) şi (IV.8,5) putem scrie

$$\sum_{i=1}^{m}p_{ij}f(a_i)=\sum_{i=1}^{m}\sum_{k=0}^{m-1}\sum_{r=0}^{m-1}\sum_{s=0}^{n-1}A_k\varphi_k(a_i)C_{rs}p_{i.}p_{.j}\varphi_r(a_i)\psi_s(b_j)=$$

$$=\sum_{r=0}^{m-1}\sum_{s=0}^{n-1}C_{rs}\sum_{k=0}^{m-1}A_k\sum_{i=0}^{m-1}p_{i.}\varphi_k(a_i)\varphi_r(a_i)\psi_s(b_j)p_{.j}.$$

Observînd că, pentru $k\neq r$, toţi coeficienţii C_{rs} sînt nuli obţinem

$$\sum_{i=1}^{m}p_{ij}f(a_i)=\sum_{r=0}^{m-1}\sum_{s=0}^{n-1}C_{rs}A_r\psi_s(b_j)p_{.j}=\sum_{r=1}^{m-1}\sum_{s=1}^{n-1}A_rC_{rs}\psi_s(b_j)p_{.j},\tag{IV.8,7}$$

deoarece $A_0=C_{0s}=C_{r0}=0$.

In acelaşi mod, găsim

$$\sum_{j=1}^{n} p_{ij}\, g(b_j) = \sum_{r=1}^{m-1} \sum_{s=1}^{n-1} B_s C_{rs} \varphi_r(a_i)\, p_{i.} \qquad \text{(IV.8,8)}$$

Pe de altă parte, din relaţiile (IV.7,5) şi (IV.7,6), stabilite în paragraful precedent, deducem

$$rp_{i.}\, f(a_i) = \sum_{j=1}^{n} p_{ij}\, g(b_j),$$

$$rp_{.j}\, g(b_j) = \sum_{i=1}^{n} p_{ij}\, f(a_i),$$

care, în baza relaţiilor (IV.8,3), (IV.8,4), (IV.8,7) şi (IV.8,8) sînt echivalente cu

$$rp_{i.} \sum_{k=1}^{m-1} A_k \varphi_k(a_i) = \sum_{r=1}^{m-1} \sum_{s=1}^{n-1} B_s C_{rs} \varphi_r(a_i)\, p_{i.},$$

$$rp_{.j} \sum_{l=1}^{n-1} B_l \psi_l(b_j) = \sum_{r=1}^{m-1} \sum_{s=1}^{n-1} A_r C_{rs} \psi_s(b_j)\, p_{.j}.$$

Aducîndu-le la forma simplă de mai jos, obţinem sistemul

$$rA_k = \sum_{s=1}^{n-1} B_s C_{ks} \qquad (k = 1, 2, \ldots, m-1),$$

$$rB_l = \sum_{r=1}^{m-1} A_r C_{rl} \qquad (l = 1, 2, \ldots, n-1),$$

de unde deducem

$$rA_k = \frac{1}{r} \sum_{s=1}^{n-1} \sum_{r=1}^{m-1} A_r C_{ks} C_{rs}$$

sau

$$r^2 A_k = \sum_{r=1}^{m-1} h_{kr} A_r \qquad (k = 1, 2, \ldots, m-1),$$

dacă notăm

$$h_{kr} = \sum_{s=1}^{n-1} C_{ks} C_{rs}.$$

Am obținut, astfel, ecuația

$$|h_{ij} - \lambda^2 \delta_{ij}| = 0,$$

care ne va da coeficientul maxim de corelație.

Este de observat că suma se extinde de la 1 pînă la m și că nu mai avem de-a face cu rădăcina banală $\lambda^2 = 1$.

Capitolul V

Distribuții cu mai multe dimensiuni

§ 1. Definirea distribuțiilor

Fie în spațiul euclidian R_s cu s dimensiuni, un vector

$$x(x_1, x_2, \ldots, x_s),$$

$x_i (i = 1, 2, \ldots, s)$ fiind componentele sale.

Fie, de asemenea, $f(x_1, x_2, \ldots, x_s)$ o funcție măsurabilă Borel și E o mulțime boreliană din spațiul R_s.

Vom spune că $f(x_1, x_2, \ldots, x_s)$ este funcția de frecvență a unei distribuții continue cu s dimensiuni, dacă

$$f(x_1, x_2, \ldots, x_s) \geqq 0$$

și

$$\iint \ldots \int_E f(x_1, x_2, \ldots, x_s)\, dx_1\, dx_2 \ldots dx_s = 1.$$

Prin *probabilitatea elementară*

$$f(x_1, x_2, \ldots, x_s)\, \Delta x_1 \Delta x_2 \ldots \Delta x_s$$

se înțelege probabilitatea ca componentele $x_1, x_2, \ldots, x_s$ ale vectorului x să fie cuprinse în domeniul elementar

$$(x_1, x_1 + \Delta x_1;\ x_2, x_2 + \Delta x_2;\ \ldots;\ x_s, x_s + \Delta x_s).$$

Scanned with OKEN Scanner

Am definit, astfel, o distribuție continuă cu s dimensiuni, a cărei funcție de frecvență este egală cu

$$\begin{cases} f(x_1, x_2, \ldots, x_s), & x \in E \\ 0, & x \in \overline{E}. \end{cases}$$

$\overline{E}$ fiind complementara lui E

Pe lîngă distribuțiile continue se întîlnesc și distribuții discrete multidimensionale, caracterizate prin faptul că vectorul $x(x_1, x_2, \ldots, x_s)$ poate să ia un număr finit sau o mulțime numărabilă de valori cu probabilități bine determinate.

Avem astfel un vector x ale cărui componente $x_1, x_2, \ldots, x_s$ pot lua respectiv valorile

$$a_{j_1}, a_{j_2}, \ldots, a_{j_s},$$

unde probabilitatea

$$p_{j_1 j_2 \ldots j_s},$$

unde

$$j_1 \in I_1, \quad j_2 \in I_2, \ldots, j_s \in I_s,$$

ar I_i $(i = 1, 2, \ldots, s)$ sînt mulțimi numărabile de indici.

Distribuția discretă cu s dimensiuni este finită, dacă fiecare dintre mulțimile de indici I_i este finită.

O distribuție oarecare este definită prin *funcția* sa de *repartiție*:

$$F(a_1, a_2, \ldots, a_s) = P(x_1 \leq a_1, x_2 \leq a_2, \ldots, x_s \leq a_s) =$$

$$= \iint \ldots \int_D dF(x_1, x_2, \ldots, x_s),$$

domeniul D fiind definit prin inegalitățile

$$x_1 \leq a_1, \quad x_2 \leq a_2, \ldots, \quad x_s \leq a_s.$$

Avem o *densitate de probabilitate* dacă

$$\frac{\partial^s F(x_1, x_2, \ldots, x_s)}{\partial x_1 \partial x_2 \ldots \partial x_s} = f(x_1, x_2, \ldots, x_s).$$

Să efectuăm transformarea

$$x_1 = g_1(y_1, y_2, \ldots, y_s),$$
$$x_2 = g_2(y_1, y_2, \ldots, y_s),$$
$$\ldots\ldots\ldots\ldots$$
$$x_s = g_s(y_1, y_2, \ldots, y_s),$$

prin care facem să corespundă vectorului $x(x_1, x_2, \ldots, x_s)$, vectorul $y(y_1, y_2, \ldots, y_s)$ şi, invers, vectorului $y(y_1, y_2, \ldots, y_s)$, vectorul $x(x_1, x_2, \ldots, x_s)$, în condiţiile care urmează.

Presupunem că funcţiile $g_i(y_1, y_2, \ldots, y_s)$ $(i = 1, 2, \ldots, s)$ sînt continue şi au derivate parţiale în raport cu y_k $(k = 1, 2, \ldots, s)$ de asemenea continue.

Dacă determinantul funcţional

$$\frac{D(x_1, x_2, \ldots, x_s)}{D(y_1, y_2, \ldots, y_s)} \neq 0,$$

în baza transformării efectuate, probabilitatea elementară de mai sus devine

$$f(x_1, x_2, \ldots, x_s)\, dx_1 dx_2 \ldots dx_s =$$
$$= f(g_1, g_2, \ldots, g_s) \left| \frac{D(x_1, x_2, \ldots, x_s)}{D(y_1, y_2, \ldots, y_s)} \right| dy_1 dy_2 \ldots dy_s.$$

§ 2. Distribuţii marginale. Distribuţii condiţionate

Vom numi *distribuţie marginală* a variabilei x_k, distribuţia variabilei x_k, indiferent de valorile luate de celelalte variabile $x_1, x_2, \ldots, x_{k-1}, x_{k+1}, \ldots, x_s$, componente ale vectorului $x(x_1, x_2, \ldots, x_s)$.

In acest caz, funcţia de repartiţie, dacă distribuţia este continuă, se va scrie

$$F_k(y) = P(x_k \leqslant y) = F(y_1, y_2, \ldots, y_s), \quad \begin{cases} y_i = +\infty, & i \neq k, \\ y_k = y. \end{cases}$$

De unde rezultă, pentru probabilitatea elementară marginală a lui x_k, dacă există o densitate de probabilitate $f_k(x_k)$,

$$f_k(x_k)\,\Delta x_k = \Delta x_k \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\dots\int_{-\infty}^{+\infty} f(x_1, x_2, \dots, x_s) \prod_{\substack{i=1\\ i\neq k}}^{s} dx_i.$$

Mai general, vom numi *distribuţie marginală a grupului de variabile*

$$x_1, x_2, \dots, x_r, \qquad (r < s)$$

distribuţia acestui grup de variabile, indiferent de valorile luate de restul variabilelor

$$x_{r+1}, x_{r+2}, \dots, x_s.$$

Cu notaţiile de mai sus, funcţia de repartiţie a acestei distribuţii marginale este

$$F_{12,\dots,r}(y_1, y_2, \dots, y_r) = F(y_1, y_2, \dots, y_r, +\infty, \dots, +\infty),$$

iar densitatea de probabilitate marginală în cazul continuu este

$$f_{12,\dots,r}(x_1, x_2, \dots, x_r) =$$

$$= \int_{-\infty}^{+\infty}\dots\int_{-\infty}^{+\infty} f(x_1, x_2, \dots, x_s)\, dx_{r+1}\, dx_{r+2}\dots dx_s.$$

Pentru determinarea *distribuţiei grupului de variabile* $x_1, x_2, \dots, x_r$, *condiţionată de valorile*

$$x_{r+1} = \xi_{r+1}, \quad x_{r+2} = \xi_{r+2}, \dots, \quad x_s = \xi_s,$$

considerăm numai cazul continuu. Notăm

$$g_{12\dots r}(x_1, x_2, \dots, x_r \mid \xi_{r+1}, \xi_{r+2}, \dots, \xi_s)$$

funcţia de frecvenţă a grupului de variabile $x_1, x_2, \dots, x_r$, dacă $x_{r+1} = \xi_{r+1}$, $x_{r+2} = \xi_{r+2}, \dots, x_s = \xi_s$.

Avem, evident,

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\dots\int_{-\infty}^{+\infty} g_{1,2,\dots,r}(x_1, x_2, \dots, x_r \mid \xi_{r+1}, \xi_{r+2}, \dots, \xi_s)\, dx_1 dx_2 \dots dx_r = 1.$$

Observăm că

$$P(y_1 < x_1 < y_1 + dy_1, \ldots, y_r < x_r < y_r + dy_r, \xi_{r+1} < x_{r+1} < \xi_{r+1} + d\xi_{r+1}, \ldots, \xi_s < x_s < \xi_s + d\xi_s) =$$

$$= P(\xi_{r+1} < x_{r+1} < \xi_{r+1} + d\xi_{r+1}, \ldots, \xi_s < x_s < \xi_s + d\xi_s) \times$$

$$\times P(y_1 < x_1 < y_1 + dy_1, \ldots, y_r < x_r < y_r + dy_r \mid \xi_{r+1} < x_{r+1} < \xi_{r+1} + d\xi_{r+1}, \ldots, \xi_s < x_s < \xi_s + d\xi_s)$$

și, prin urmare,

$$g_{12\ldots r}(y_1, y_2, \ldots, y_r \mid \xi_{r+1}, \xi_{r+2}, \ldots, \xi_s) = \frac{f(y_1, y_2, \ldots, y_r, \xi_{r+1}, \ldots, \xi_s)}{\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty} f(y_1, y_2, \ldots, y_r, \xi_{r+1}, \ldots, \xi_s)\, d\xi_{r+1}\ldots d\xi_s}.$$

§ 3. Momente

Momentele de ordinul întîi sînt definite de integralele de tip Stieltjes

$$m_k = \iint\ldots\int_{R_s} x_k dF(x_1, x_2, \ldots, x_s) \qquad (k = 1, 2, \ldots, s)$$

dacă ele există. Momentele de ordinul al doilea se definesc prin integralele

$$M_{ij} = \iint\ldots\int_{R_s} x_i x_j dF(x_1, x_2, \ldots, x_s) \qquad (i, j = 1, 2, \ldots, s).$$

În general,

$$M_{r_1, r_2 \ldots r_s} = \iint\ldots\int_{R_s} x_1^{r_1} x_2^{r_2} \ldots x_s^{r_s}\, dF(x_1, x_2, \ldots, x_s),$$

de asemenea dacă ele există.

In ceea ce priveşte momentele centrate, dacă ele există, se exprimă cum urmează

$$m_{ij} = \iint \ldots \int_{R_s} (x_i - m_i)(x_j - m_j)\, dF(x_1, x_2, \ldots, x_s),$$

$$m_{r_1 r_2 \ldots r_s} = \iint \ldots \int_{R_s} (x_1 - m_1)^{r_1} (x_2 - m_2)^{r_2} \ldots (x_s - m_s)^{r_s}\, dF(x_1, x_2, \ldots, x_r).$$

După cum se observă, momentele centrate ale distribuţiei cu mai multe dimensiuni se pot exprima cu ajutorul momentelor ei obişnuite. Prin definiţie, *funcţia caracteristică a unei distribuţii cu s dimensiuni* este dată de expresia

$$\varphi(t_1, t_2, \ldots, t_s) = M(e^{it'x}) = \iint \ldots \int_{R_s} e^{it'x}\, dF(x_1, x_2, \ldots, x_s),$$

unde $t' = \| t_1, t_2, \ldots, t_s \|$ este transpusa vectorului coloană

$$t = \left\| \begin{matrix} t_1 \\ t_2 \\ \vdots \\ t_s \end{matrix} \right\|,$$

unde ţinem seama că produsul scalar $t'x$ este

$$t_1 x_1 + t_2 x_2 + \ldots + t_s x_s.$$

Cunoscînd funcţia caracteristică a unei distribuţii cu s dimensiuni, momentele ei de diverse ordine, care există, sînt date de derivatele acestei funcţii în punctul $t_1 = t_2 = \ldots = t_s = 0$

$$m_k = \frac{1}{i} \left[\frac{\partial \varphi(t_1, t_2, \ldots, t_s)}{\partial t_k} \right]_{t_1, t_2, \ldots, t_s = 0},$$

$$M_{ij} = \frac{1}{i^2} \left[\frac{\partial^2 \varphi(t_1, t_2, \ldots, t_s)}{\partial t_i \cdot \partial t_j} \right]_{t_1, t_2, \ldots, t_s = 0},$$

. . . . . .

$$M_{r_1 r_2 \ldots r_s} = \frac{1}{i^{r_1 + r_2 + \ldots + r_s}} \left[\frac{\partial^{r_1 + r_2 + \ldots + r_s} \varphi(t_1, t_2, \ldots, t_s)}{\partial t_1^{r_1} \partial t_2^{r_2} \ldots \partial t_s^{r_s}} \right]_{t_1, t_2, \ldots, t_s = 0}.$$

Observăm că derivatele în punctul $t=0$ există o dată cu momentele respective.

Valoarea medie a produsului

$$x_1^{\lambda_1} x_2^{\lambda_2} \dots x_r^{\lambda_r},$$

condiționată de valorile

$$x_{r+1}=\xi_{r+1}, \quad x_{r+2}=\xi_{r+2}, \dots, \quad x_s=\xi_s$$

se notează cu

$$M_{\lambda_1\lambda_2\dots\lambda_r}(x_1, x_2, \dots, x_r \mid \xi_{r+1}, \dots, \xi_s).$$

Presupunînd că distribuția este continuă vom avea

$$M_{\lambda_1\lambda_2\dots\lambda_r}(x_1, x_2, \dots, x_r \mid \xi_{r+1}, \dots, \xi_s) =$$

$$= \iint \dots \int_{R_s} x_1^{\lambda_1} x_2^{\lambda_2} \dots x_r^{\lambda_r} g_{12\dots r}(x_1, x_2, \dots, x_r \mid \xi_{r+1}, \dots, \xi_s)\, dx_1 \dots dx_s,$$

unde

$$g_{12\dots r}(x_1, x_2, \dots, x_r \mid \xi_{r+1}, \dots, \xi_s) =$$

$$= \frac{f(x_1, x_2, \dots, x_r, \xi_{r+1}, \dots, \xi_s)}{\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} f(x_1, x_2, \dots, x_r, \xi_{r+1}, \dots, \xi_s)\, d\xi_{r+1} \dots d\xi_s}$$

este funcția de frecvență a distribuției grupului de variabile $x_1, x_2, \dots, x_r$, *condiționată* de valorile grupului de variabile x_{r+1}, $x_{r+2}, \dots, x_s$.

§ 4. Regresiune. Corelație parțială

Să notăm valoarea medie a variabilei x_1, condiționată de egalitățile

$$x_2=\xi_2, \quad x_3=\xi_3, \dots, \quad x_s=\xi_s$$

cu

$$m_{1\cdot\xi_2\xi_3\dots\xi_s}.$$

Atunci, în cazul continuu, vom avea

$$m_{1\cdot\xi_2\xi_3\dots\xi_s} = M(x_1 \mid \xi_2\xi_3 \dots \xi_s) = \int_{-\infty}^{+\infty} x_1 g_1(x_1 \mid \xi_2 \dots \xi_s)\, dx_1,$$

Scanned with OKEN Scanner

unde

$$g_1(x_1 | \xi_2 \dots \xi_s) = \frac{f(x_1; \xi_2, \dots, \xi_s)}{\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} f(x_1; \xi_2 \dots \xi_s)\, d\xi_2 \dots d\xi_s}.$$

După cum se vede, media condiţionată $m_{1\xi_2 \dots \xi_s}$ este o funcţie de $\xi_2, \xi_3, \dots, \xi_s$.

Hipersuprafaţa a cărei ecuaţie este

$$x_1 = m_{1x_2x_3 \dots x_s}$$

poartă numele de *suprafaţă de regresiune a variabilelor* $x_2, \dots, x_s$ *faţă de* x_1. Dacă această hipersuprafaţă este un hiperplan, avem o *regresiune liniară a variabilelor* $x_2, x_3, \dots, x_s$ *asupra lui* x_1.

In mod analog definim regresiunea variabilelor

$$x_1, x_2, \dots, x_{k-1}, x_{k+1}, \dots, x_s$$

asupra lui x_k.

Expresia

$$\sigma_{1 \cdot \xi_2 \xi_3 \dots \xi_s} = \int_{-\infty}^{+\infty} (x_1 - m_{1\xi_2\xi_3 \dots \xi_s})^2 g_1(x_1 | \xi_2 \dots \xi_s)\, dx_1$$

este *dispersia distribuţiei cu s dimensiuni în raport cu regresiunea considerată.*

Să introducem următoarele notaţii pentru diversele valori medii condiţionate care urmează

$$m_{1 \cdot 34 \dots s} = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} x_1 g_{12}(x_1, x_2 | \xi_3, \dots, \xi_s)\, dx_1\, dx_2,$$

$$m_{2 \cdot 34 \dots s} = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} x_2 g_{12}(x_1, x_2 | \xi_3, \dots, \xi_s)\, dx_1\, dx_2,$$

$$\sigma^2_{1 \cdot 34 \dots s} = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x_1 - m_{1 \cdot 34 \dots s})^2 g_{12}(x_1, x_2 | \xi_3, \dots, \xi_s)\, dx_1\, dx_2,$$

$$\sigma^2_{2 \cdot 34 \dots s} = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x_2 - m_{2 \cdot 34 \dots s})^2 g_{12}(x_1, x_2 | \xi_3, \dots, \xi_s)\, dx_1\, dx_2.$$

Scanned with OKEN Scanner

Coeficientul $\rho_{12\ldots s}$, definit de expresia

$$\rho_{12\ldots s}=\frac{\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(x_1-m_{1\cdot 34\ldots s})^2(x_2-m_{2\cdot 34\ldots s})^2 g_{12}(x_1, x_2 \mid \xi_3,\ldots,\xi_s)\,dx_1 dx_2}{\sigma_{1\cdot 34\ldots s}\cdot\sigma_{2\cdot 34\ldots s}}$$

se numeşte *coeficient de corelaţie parţială al celor două variabile* x_1 şi x_2, *faţă de* $x_3, x_4, \ldots, x_s$. El este, după această definiţie, o funcţie de variabilele $x_3, x_4, \ldots, x_s$.

§ 5. Rangul unei distribuţii

O distribuţie într-un spaţiu euclidian cu s dimensiuni R_s este de rangul $r \leq s$, dacă toate punctele ei se găsesc într-o varietate liniară V_r cu r dimensiuni, cuprinsă în spaţiul R_s, şi nu se găseşte, fără a fi cuprinsă în acest spaţiu, într-o varietate liniară cu un grad mai mic decît r.

De exemplu, în spaţiul euclidian cu trei dimensiuni R_3, o distribuţie de rangul întîi are toate punctele ei situate pe o linie dreaptă fără să se reducă la un punct unic, o distribuţie de rangul al doilea are toate punctele ei situate într-un plan, fără să existe vreo dreaptă care să le cuprindă pe toate, iar o distribuţie de rangul al treilea va avea toate punctele ei aşezate în spaţiu, fără să existe vreun plan pe care să se găsească toate punctele distribuţiei.

Revenind la spaţiul euclidian cu s dimensiuni, să considerăm următoarea formă pătratică definită pozitivă.

$$M[\lambda'(x-m)]^2=\sum_{i=1}^{s}\sum_{j=1}^{s} m_{ij}\lambda_i\lambda_j=\lambda' M\lambda \geq 0,$$

unde $\lambda'=\|\lambda_1, \lambda_2, \ldots, \lambda_s\|$ este transpusul vectorului coloană

$$\lambda=\left\|\begin{matrix}\lambda_1\\ \lambda_2\\ \vdots\\ \lambda_s\end{matrix}\right\|.$$

Se ştie că, dacă matricea covariantă $M=\|m_{ij}\|$ este de rangul s determinantul $|m_{ij}|>0$. Dacă, însă, matricea $\|m_{ij}\|$ este de rangul

$s-1$, urmează că există o valoare particulară a vectorului $\lambda(\lambda_1, \lambda_2, \ldots, \lambda_s)$, care verifică egalitatea

$$\mu'(x-m)=\mu_1(x_1-m_1)+\mu_2(x_2-m_2)+\ldots+\mu_s(x_s-m_s)=0,$$

μ' fiind transpusa vectorului μ. In acest caz, toate punctele distribuţiei s-dimensionale se vor găsi pe un hiperplan a cărui ecuaţie este reprezentată de însăşi egalitatea de mai sus, deci rangul distribuţiei considerate va fi $s-1$.

Presupunînd că rangul matricei $\|m_{ij}\|$ este egal cu $s-k$, vom găsi k valori particulare ale vectorului $\lambda(\lambda_1, \lambda_2, \ldots, \lambda_s)$

$$\mu_1, \mu_2, \ldots, \mu_k$$

astfel ca să aibă loc egalităţile

$$\mu_1'(x-m)=0, \mu_2'(x-m)=0, \ldots, \mu_k'(x-m)=0.$$

Inseamnă că distribuţia se găseşte în fiecare dintre aceste hiperplane ale căror ecuaţii sînt exprimate prin egalităţile mai sus stabilite şi, deci, rangul distribuţiei va fi egal cu $s-k$.

In concluzie, putem spune că rangul unei distribuţii într-un spaţiu euclidian cu mai multe dimensiuni este dat de rangul matricei covariante.

La aceste rezultate am ajuns plecînd de la forma pătratică definită pozitivă

$$\int\int\cdots\int_{R_s}[\lambda_1(x_1-m_1)+\lambda_2(x_2-m_2)+\ldots$$
$$\ldots+\lambda_s(x_s-m_s)]^2\,dF(x_1, x_2, \ldots, x_s)\geqq 0.$$

Să considerăm şi forma pătratică pozitiv definită de $\lambda_r, \lambda_{r+1}, \ldots, \lambda_s$

$$\int\int\cdots\int_{R_s}[\lambda_r(x_r-m_r)+\lambda_{r+1}(x_{r+1}-m_{r+1})+\ldots$$
$$\ldots+\lambda_s(x_s-m_s)]^2\,dF(x_1, x_2, \ldots, x_s)\geqq 0.$$

Discriminantul ei trebuie să fie pozitiv, adică

$$\begin{vmatrix} m_{r,r} & m_{r,r+1} & \ldots & m_{r,s} \\ m_{r+1,r} & m_{r+1,r+2} & \ldots & m_{r+1,s} \\ \cdot & \cdot & \cdot & \cdot \\ m_{s,r} & m_{s,r+1} & \ldots & m_{s,s} \end{vmatrix} \geqq 0.$$

Discriminantul în cauză va fi nul dacă există un sistem de valori, pentru care

$$\mu_r (x_r - m_r) + \mu_{r+1} (x_{r+1} - m_{r+1}) + \ldots + \mu_s (x_s - m_s) = 0.$$

După cum se ştie, în acest caz, toate punctele distribuţiei se vor găsi pe un hiperplan a cărui ecuaţie este reprezentată de această egalitate şi, deci, rangul distribuţiei se micşorează cu o unitate.
In urma celor de mai sus, rezultă că dacă rangul distribuţiei este s, toţi determinanţii de forma celui de mai sus vor fi strict pozitivi.

§ 6. Funcţii de regresiune

Se pot determina funcţii de regresiune pentru distribuţii cu mai multe dimenisiuni, folosind hipersuprafeţele de ajustare, care decurg din metoda celor mai mici pătrate.
Considerînd funcţia liniară de variabile $x_2, x_3, \ldots, x_s$

$$\alpha_1 + \alpha_2 x_2 + \alpha_3 x_3 + \ldots + \alpha_s x_s,$$

ne propunem să determinăm parametrii

$$\alpha_1, \alpha_2, \ldots, \alpha_s$$

astfel ca expresia

$$M (x_1 - \alpha_1 - \alpha_2 x_2 - \ldots - \alpha_s x_s)^2$$

să fie minimă.
Presupunînd cazul continuu, aceasta revine a face minimă integrala

$$\iint \cdots \int_{R_s} (x_1 - \alpha_1 - \alpha_2 x_2 - \ldots - \alpha_s x_s)^2 f(x_1, x_2, \ldots, x_s)\, dx_1\, dx_2 \ldots dx_s.$$

Anulînd derivata ei parţială de ordinul întîi în raport cu α_1, vom avea cu notaţiile adoptate

$$\alpha_1 + m_2 \alpha_2 + \ldots + m_s \alpha_s = m_1.$$

De unde deducem

$$\alpha_1 = m_1 - m_2 \alpha_2 - \ldots - m_s \alpha_s.$$

Atunci, integrala de mai sus devine

$$\iint \cdots \int_{R_s} [x_1 - m_1 - \alpha_2 (x_2 - m_2) - \ldots$$

$$\ldots - \alpha_s (x_s - m_s)]^2 f(x_1, x_2, \ldots, x_s)\, dx_1\, dx_2 \ldots dx_s.$$

Scanned with OKEN Scanner

Anulînd, de asemenea, derivatele ei parţiale de ordinul întîi în raport cu $\alpha_2, \alpha_3, \ldots, \alpha_s$, vom găsi sistemul liniar de mai jos cu $s-1$ necunoscute $(\alpha_2, \alpha_3, \ldots, \alpha_s)$:

$$\begin{aligned}
&m_{22}\alpha_2 + m_{23}\alpha_3 + \ldots + m_{2s}\alpha_s = m_{12},\\
&m_{32}\alpha_2 + m_{33}\alpha_3 + \ldots + m_{3s}\alpha_s = m_{13},\\
&\cdots\cdots\cdots\cdots\cdots\cdots\\
&m_{s2}\alpha_2 + m_{s3}\alpha_3 + \ldots + m_{ss}\alpha_s = m_{1s}.
\end{aligned}$$

Acest sistem admite soluţii unic determinate, deoarece matricea

$$A = \begin{Vmatrix} m_{22} & m_{23} \ldots m_{2s} \\ m_{32} & m_{33} \ldots m_{3s} \\ \cdots & \cdots\cdots \\ m_{s2} & m_{s3} \ldots m_{ss} \end{Vmatrix},$$

fiind de rangul $s-1$, determinantul sistemului, adică $|A|$, este diferit de zero.

Vom obţine, astfel, următoarele soluţii ale sistemului de mai sus

$$\alpha_k = \frac{\sum_{i=2}^{s} m_{1i} A_{ik}}{|A|} \qquad (k = 2, 3, \ldots, s),$$

unde am notat cu A_{ij} complementul algebric al elementului m_{ij} din determinantul $|A|$.

Se poate acum determina şi parametrul α_1 din relaţia

$$\alpha_1 = m_1 - m_2\alpha_2 - \ldots - m_s\alpha_s.$$

Am determinat, deci, parametrii $\alpha_1, \alpha_2, \ldots, \alpha_s$, pentru care integrala

$$\iint \cdots \int_{R_s} [x - m_1 - \alpha_2(x_2 - m_2) - \ldots$$

$$\ldots - \alpha_s(x_s - m_s)]^2 f(x_1, x_2, \ldots, x_s)\, dx_1\, dx_2 \ldots dx_s$$

este minimă.

Minimul acestei integrale este dat de expresia

$$\iint\cdots\int_{R_s}\left[x_1-m_1-\frac{1}{|A|}\sum_{i=2}^{s}\sum_{k=2}^{s}m_{1i}\times\right.$$

$$\left.\times A_{ik}(x_k-m_k)\right]^2 f(x_1,\ x_2,\ \ldots,\ x_s)\,dx_1\,dx_2\ldots dx_s$$

care, conform cu notațiile fixate, se mai scrie

$$m_{11}-\frac{2}{|A|}\sum_{i=2}^{s}\sum_{k=2}^{s}m_{1i}m_{1k}A_{ik}+$$

$$+\frac{1}{|A|^2}\sum_{i=2}^{s}\sum_{k=2}^{s}\sum_{i'=2}^{s}\sum_{k'=2}^{s}m_{1i}m_{1i'}A_{ik}A_{i'k'}m_{kk'}=$$

$$=m_{11}-\frac{2}{|A|}\sum_{i=2}^{s}\sum_{k=2}^{s}m_{1i}m_{1k}A_{ik}+$$

$$+\frac{1}{|A|^2}\sum_{i=2}^{s}\sum_{k=2}^{s}\sum_{i'=2}^{s}m_{1i}m_{1i'}A_{ik}\sum_{k'=2}^{s}A_{i'k'}m_{kk'}.$$

Observăm că pentru $i'\neq k$,

$$\sum_{k'=2}^{s}A_{i'k'}m_{kk'}=0,$$

ca sumă a produselor elementelor liniei k prin complementții algebrici ai elementelor liniei i', iar pentru $i'=k$, avem

$$\sum_{k'=2}^{s}A_{i'k'}m_{kk'}=|A|.$$

Atunci, minimul căutat este

$$m_{11}-\frac{2}{|A|}\sum_{i=2}^{s}\sum_{k=2}^{s}m_{1i}\cdot m_{1k}A_{ik}+\frac{1}{|A|}\sum_{i=2}^{s}\sum_{k=2}^{s}m_{1i}\cdot m_{1k}A_{ik}=$$

$$=m_{11}-\frac{1}{|A|}\sum_{i=2}^{s}\sum_{k=2}^{s}m_{1i}\cdot m_{1k}\cdot A_{ik}=\frac{m_{11}|A|-\sum_{i=2}^{s}\sum_{k=2}^{s}m_{1i}\cdot m_{1k}\cdot A_{ik}}{|A|}=$$

$$=\frac{m_{11}|A|-\sum_{i=2}^{s}m_{1i}\sum_{k=2}^{s}m_{1k}A_{ik}}{|A|}=\frac{m_{11}|A|-\sum_{i=2}^{s}m_{1i}|A|}{|A|},$$

deoarece, în baza observaţiei de mai sus, $\sum_{k=2}^{s} m_{1k}A_{ik}=0$, dacă $i\neq 1$ şi $\sum_{k=2}^{s} m_{1k}A_{ik}=|A|$, dacă $i=1$.

Considerînd matricea de rangul s

$$M=\begin{Vmatrix} m_{11} & m_{12}\ldots m_{1s} \\ m_{21} & m_{22}\ldots m_{2s} \\ \cdot & \cdot\ \cdot\ \cdot\ \cdot\ \cdot \\ m_{s1} & m_{s2}\ldots m_{ss} \end{Vmatrix},$$

se vede că $|A|$ este minorul elementului m_{11} din determinantul $|M|$. Insemnînd prin M_{ij} complementul algebric al elementului m_{ij} din $|M|$, avem evident

$$|A|=M_{11}.$$

Prin urmare, minimul de mai sus devine

$$\frac{m_{11}M_{11}-\sum_{i=2}^{s} m_{1i}M_{1i}}{|A|}=\frac{m_{11}M_{11}+\sum_{i=2}^{s} m_{1i}M_{1i}}{|A|}=\frac{|M|}{|A|},$$

ţinînd seama că, prin permutarea coloanelor *1* şi *i* între ele, în determinantul $|A|$, rezultă

$$M_{i1}=-M_{1i}.$$

Prin urmare

$$x_1-m_1=\frac{1}{|A|}\sum_{i=2}^{s}\sum_{k=2}^{s} m_{1i}A_{ik}(x_k-m_k)$$

reprezintă *hiperplanul de regresiune al variabilelor* $x_2, x_3, \ldots, x_s$ *asupra variabilei* x_1.

După cum se vede, el trece prin centrul de greutate $G(m_1, m_2, \ldots, m_s)$ al distribuţiei şi se bucură de proprietatea remarcabilă că valoarea medie a pătratelor distanţelor d, de la orice punct $P(x_1, x_2, \ldots, x_s)$ al distribuţiei date la acest hiperplan, luate paralel cu axa Ox_1, este minimă, avînd

$$M(d^2)=\frac{|M|}{|A|}.$$

Scanned with OKEN Scanner

Din acest punct de vedere, înseamnă că hiperplanul considerat este dintre toate hiperplanele din R_s cel mai apropiat în direcţia axei Ox_1 de punctele distribuţiei.

Dacă minimul $\frac{|M|}{|A|}$ este nul, ceea ce are loc cînd $|M|=0$ şi $|A|\neq 0$, urmează că toate punctele distribuţiei date sînt aşezate pe hiperplanul de regresiune al variabilelor $x_2, x_3, \ldots, x_s$ faţă de variabila x_1.

Există s hiperplane de regresiune, care trec toate prin centrul de greutate al distribuţiei şi se bucură de proprietatea arătată. In cazul cînd minimul este egal cu zero, toate cele s hiperplane de regresiune vor coincide şi punctele distribuţiei se vor găsi numai într-un singur hiperplan de regresiune.

In particular, pentru $s=2$, ecuaţia hiperplanului de regresiune al variabilelor $x_2, x_3, \ldots, x_s$ asupra variabilei x_1 devine

$$x_1-m_1=\frac{m_{12}}{m_{22}}(x_2-m_2)$$

sau

$$x_1-m_1=\frac{\rho\sigma_1}{\sigma_2}(x_2-m_2).$$

In mod analog, ecuaţia hiperplanului de regresiune al variabilelor $x_1, x_3, \ldots, x_s$ asupra variabilei x_2 se reduce la ecuaţia

$$x_2-m_2=\frac{\rho\sigma_2}{\sigma_1}(x_1-m_1).$$

Am regăsit, astfel, ecuaţiile dreptelor de regresiune stabilite la studiul distribuţiilor de două variabile.

In ceea ce priveşte minimul, el devine

$$\frac{|M|}{|A|}=\frac{\begin{vmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{vmatrix}}{m_{22}}=\frac{\begin{vmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{vmatrix}}{\sigma_2^2}=\sigma_1^2(1-\rho^2),$$

respectiv $\sigma_2^2(1-\rho^2)$, după cum considerăm primul sau al doilea hiperplan de regresiune.

§ 7. Coeficientul multiplu de corelaţie

Funcţia

$$F_1(x_2,\ x_3,\ldots,x_s)=\frac{1}{|A|}\sum_{i=2}^{s}\sum_{k=2}^{s}m_{1i}A_{ik}(x_k-m_k)$$

poartă numele de *funcţie de regresiune referitoare la variabila* x_1.
De aici urmează că există s funcţii de regresiune.
Coeficientul de corelaţie între variabilele

$$x_1-m_1,\ F_1(x_2,\ x_3,\ldots,x_s),$$

pe care-l vom nota cu $R_{1\cdot 23\ldots s}$ se numeşte *coeficient multiplu de corelaţie.*
Atunci, în baza definiţiei sale, coeficientul multiplu de corelaţie este

$$R_{1\cdot 23\ldots s}=\frac{M[(x_1-m_1)\cdot F_1(x_2,\ x_3,\ldots,x_s)]}{\sqrt{M(x_1-m_1)^2}\cdot\sqrt{M[F_1(x_2,\ x_3,\ldots,x_s)]^2}}.$$

Procedînd la calculul numărătorului acestei expresii, găsim

$$M[(x_1-m_1)F_1(x_2,\ x_3,\ldots,x_s)]=$$

$$=M[(x_1-m_1)]\frac{1}{|A|}\sum_{i=2}^{s}\sum_{k=2}^{s}m_{1i}A_{ik}(x_k-m_k)=$$

$$=\frac{1}{|A|}\sum_{i=2}^{s}\sum_{k=2}^{s}m_{1i}A_{ik}M[(x_1-m_1)(x_k-m_k)]=\frac{1}{|A|}\sum_{i=2}^{s}\sum_{k=2}^{s}m_{1i}m_{ik}A_{ik}.$$

Pentru numitor

$$M(x-m_1)^2=m_{11},$$

iar

$$M[F_1(x_2,\ x_3,\ldots,x_s)]^2=\frac{1}{|A|^2}M\Big[\sum_{i=2}^{s}\sum_{k=2}^{s}m_{1i}A_{ik}(x_k-m_k)\Big]^2=$$

$$=\frac{1}{|A|}\sum_{i=2}^{s}\sum_{k=2}^{s}m_{1i}m_{1k}A_{ik},$$

potrivit celor stabilite mai înainte.

Coeficientul multiplu de corelaţie devine

$$R_{1\,.\,23\,\ldots\,s}=\frac{\sqrt{\sum_{i=2}^{s}\sum_{k=2}^{s}m_{1i}m_{1k}A_{ik}}}{\sqrt{m_{11}\,|\,A\,|}}.$$

sau ridicînd la pătrat,

$$R^2_{1\,.\,23\,\ldots\,s}=\frac{\sum_{i=2}^{s}\sum_{k=2}^{s}m_{1i}m_{1k}A_{ik}}{m_{11}\,|\,A\,|}.$$

Din relaţia cunoscută

$$m_{11}-\frac{\sum_{i=2}^{s}\sum_{k=2}^{s}m_{1i}m_{1k}A_{ik}}{|\,A\,|}=\frac{|\,M\,|}{|\,A\,|},$$

deducem

$$R^2_{1\,.\,23\,\ldots\,s}=\frac{m_{11}-\frac{|\,M\,|}{|\,A\,|}}{m_{11}}=1-\frac{|\,M\,|}{m_{11}\,|\,A\,|}\quad\left(\frac{|\,M\,|}{m_{11}\,|\,A\,|}\leqslant 1\right),$$

de unde urmează că

$$\frac{|\,M\,|}{|\,A\,|}=m_{11}\,(1-R^2_{1\,.\,23\,\ldots\,s}),$$

adică expresia minimului, corespunzătoare aceleia pe care am stabilit-o în paragraful precedent.

Dacă coeficientul multiplu de corelaţie $R_{1\,.\,23\,\ldots\,s}$ este egal cu ± 1, minimul $\frac{|\,M\,|}{|\,A\,|}$ este nul, ceea ce înseamnă că toate punctele distribuţiei date se găsesc în hiperplanul de regresiune al variabilelor $x_2, x_3, \ldots, x_s$ asupra variabilei x_1 şi, deci, distribuţia va avea $s-1$ dimensiuni, în loc de s dimensiuni.

După cum se constată, coeficientul multiplu de corelaţie păstrează aceeaşi proprietate ca şi coeficientul de corelaţie între variabilele unei distribuţii cu două dimensiuni şi anume se poate presupune, cînd el are o valoare foarte apropiată de ± 1, că distribuţia dată are toate punctele ei pe un hiperplan.

Scanned with OKEN Scanner

§ 8. Studiul regresiunii cu ajutorul regulii celor mai mici pătrate

Fiind dată o distribuţie continuă cu s dimensiuni, să determinăm în spaţiul R_s un hiperplan, astfel ca valoarea medie a pătratelor distanţelor d de la punctele distribuţiei la acest hiperplan să fie minimă. Fie $P(x_1, x_2, \ldots, x_s)$ un punct al distribuţiei date şi

$$\alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_s x_s - p = 0$$

ecuaţia hiperplanului căutat, unde $\alpha_i (i = 1, 2, \ldots, s)$ sînt cosinuşii directori ai normalei dusă din origine la plan care verifică relaţia

$$\alpha_1^2 + \alpha_2^2 + \ldots + \alpha_s^2 = 1.$$

Atunci distanţa de la punctul $P(x_1, x_2, \ldots, x_s)$ la hiperplanul definit de ecuaţia de mai sus va fi

$$\alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_s x_s - p$$

şi, prin urmare, totul revine la determinarea parametrilor $\alpha_1, \alpha_2, \ldots, \alpha_s$ şi p, verificînd relaţia precedentă şi pentru care este minimă expresia

$$M(d^2) = \iint \cdots \int_{R_s} (\alpha_1 x_1 + \alpha_2 x_2 + \ldots$$

$$\ldots + \alpha_s x_s - p)^2 f(x_1, x_2, \ldots, x_s)\, dx_1\, dx_2 \ldots dx_s.$$

Anulînd derivata acesteia în raport cu p, obţinem cu notaţiile noastre

$$\alpha_1 m_1 + \alpha_2 m_2 + \ldots + \alpha_s m_s = p,$$

de unde rezultă că

$$M(d^2) = \iint \cdots \int_{R_s} [\alpha_1 (x_1 - m_1) + \alpha_2 (x_2 - m_2) + \ldots$$

$$\ldots + \alpha_s (x_s - m_s)]^2 f(x_1, x_2, \ldots, x_s)\, dx_1\, dx_2 \ldots dx_s = min.$$

Insă aceasta, fiind o formă pătratică definită pozitivă, se poate pune sub forma

$$\sum_{i=1}^{s} \sum_{j=1}^{s} \alpha_i \alpha_j m_{ij} = \alpha' M \alpha,$$

M fiind matricea covariantă $\| m_{ij} \|$ de rangul s, iar α' transpusa vectorului $\alpha\,(\alpha_1, \alpha_2, \ldots, \alpha_s)$.
Se ştie că se poate face întotdeauna o transformare ortogonală $\alpha = H\beta$, în aşa fel încît

$$\alpha' M \alpha = \sum_{i=1}^{s} \lambda_i \beta_i^2,$$

unde λ_i sînt rădăcinile ecuaţiei de gradul s

$$|M - \lambda E| = \begin{vmatrix} m_{11}-\lambda & m_{12} & \ldots & m_{1s} \\ m_{21} & m_{22}-\lambda & \ldots & m_{2s} \\ \ldots & \ldots & \ldots & \ldots \\ m_{s1} & m_{s2} & \ldots & m_{ss}-\lambda \end{vmatrix} = 0.$$

Forma pătratică, fiind definită pozitivă, valorile proprii $\lambda_i\,(i = 1, 2, \ldots, s)$ sînt reale şi pozitive. Presupunînd că λ_h este cea mai mică dintre rădăcinile $\lambda_1, \lambda_2, \ldots, \lambda_s$, vom avea

$$\sum_{i=1}^{s} \lambda_i \beta_i^2 = \lambda_1 \beta_1^2 + \lambda_2 \beta_2^2 + \ldots + \lambda_h \beta_h^2 + \ldots + \lambda_s \beta_s^2 =$$

$$= \lambda_h (\beta_1^2 + \beta_2^2 + \ldots + \beta_s^2) + (\lambda_1 - \lambda_h)\beta_1^2 + (\lambda_2 - \lambda_h)\beta_2^2 + \ldots + (\lambda_s - \lambda_h)\beta_s^2 =$$

$$= \lambda_h \sum_{i=1}^{s} \beta_i^2 + \sum_{i=1}^{s} (\lambda_i - \lambda_h)\beta_i^2 = \lambda_h + \sum_{i=1}^{s} (\lambda_i - \lambda_h)\beta_i^2,$$

avînd în vedere că, prin transformarea ortogonală $\alpha = H\beta$, rezultă

$$\sum_{i=1}^{s} \alpha_i^2 = \sum_{i=1}^{s} \beta_i^2 = 1.$$

Observînd că $\lambda_i - \lambda_h > 0$, dacă $i \neq h$, şi $\lambda_i - \lambda_h = 0$, dacă $i = h$, expresia

$$\alpha' M \alpha = \lambda_h + \sum_{i=1}^{s} (\lambda_i - \lambda_h)\beta_i^2$$

va fi minimă cînd

$$\beta_1 = 0, \quad \beta_2 = 0, \ldots, \quad \beta_{h-1} = 0, \quad \beta_{h+1} = 0, \ldots, \quad \beta_s = 0.$$

De aici şi în baza relaţiei $\sum_{i=1}^{s} \beta_i^2 = 1$, deducem

$$\beta_h = 1,$$

Scanned with OKEN Scanner

ceea ce înseamnă că vectorului $\alpha(\alpha_1, \alpha_2, \ldots, \alpha_s)$ în urma transformării $\alpha = H\beta$, îi corespunde o singură valoare bine determinată.

Prin urmare, există un singur hiperplan astfel că valoarea medie a pătratelor distanţelor de la orice punct al distribuţiei date la acest hiperplan să fie minimă.

Minimul este egal, după cum s-a arătat mai sus, cu cea mai mică rădăcină λ_h a ecuaţiei de gradul s

$$\begin{vmatrix} m_{11}-\lambda & m_{12} \ldots m_{1s} \\ m_{21} & m_{22}-\lambda \ldots m_{2s} \\ \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \\ m_{s1} & m_{s2} \ldots m_{ss}-\lambda \end{vmatrix} = 0.$$

Să examinăm cazul cînd rădăcina cea mai mică λ_h nu este simplă. Admiţînd că există două rădăcini egale $\lambda_h = \lambda_k$, minimul formei noastre pătratice pozitiv definite

$$\alpha' M\alpha = \lambda_h + \sum_{i=1}^{s} (\lambda_i - \lambda_h)\beta_i^2$$

va avea loc, dacă

$$\beta_1 = 0, \quad \beta_2 = 0, \ldots, \beta_{h-1} = 0, \quad \beta_{h+1} = 0, \ldots, \beta_{k-1} = 0, \quad \beta_{k+1} = 0, \ldots, \beta_s = 0,$$

deoarece $\lambda_i - \lambda_h \neq 0$, pentru $i \neq h, k$ şi $\lambda_i - \lambda_h = 0$, pentru $i = h, k$.

Din relaţia $\sum_{i=1}^{s} \beta_i^2 = 1$, urmează că

$$\beta_h^2 + \beta_k^2 = 1,$$

adică, vectorului $\alpha(\alpha_1, \alpha_2, \ldots, \alpha_s)$ nu-i mai corespunde acum o valoare unic determinată, ci o infinitate de valori.

In consecinţă, dacă cea mai mică rădăcină λ_h a ecuaţiei $|M - \lambda E| = 0$ este multiplă, atunci există o infinitate de hiperplane pentru care valoarea medie a pătratelor distanţelor de la orice punct al distribuţiei considerate la fiecare dintre aceste hiperplane este minimă.

In particular, dacă toate rădăcinile $\lambda_i (i = 1, 2, \ldots, s)$ sînt egale, atunci toate hiperplanele, care trec prin centrul distribuţiei, se bucură de această proprietate.

Capitolul VI

Distribuția normală

§ 1. Considerații generale

Distribuția normală este fundamentală în statistica matematică. Ea este definită de funcția de frecvență

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-m)^2}{2\sigma^2}}, \qquad \text{(VI.1,1)}$$

unde x poate să ia toate valorile reale $(-\infty<x<+\infty)$.

Funcția $f(x)$ este funcție de frecvență deoarece, pe de o parte, ea este continuă și pozitivă pentru orice x real iar pe de altă parte, avem

$$\int_{-\infty}^{+\infty} f(x)\,dx=1. \qquad \text{(IV.1,2)}$$

Făcînd substituția $y=\frac{x-m}{\sqrt{2}\,\sigma}$ obținem

$$\int_{-\infty}^{+\infty} f(x)\,dx=\frac{1}{\sqrt{2\pi}\,\sigma}\int_{-\infty}^{+\infty} e^{-\frac{(x-m)^2}{2\sigma}}\,dx=\frac{1}{\sqrt{\pi}}\int_{-\infty}^{+\infty} e^{-y^2}dy.$$

Dar, după cum am arătat, integrala $\frac{1}{\sqrt{\pi}}\int_{-\infty}^{+\infty} e^{-y^2}dy$ are valoarea 1 și, deci, relația (VI.1,2) este stabilită.

Observăm că, funcția de frecvență (VI.1,1) a distribuției normale depinde de parametrii m și σ. Vom nota de aici înainte prin $N(m, \sigma)$ orice distribuție normală a cărei funcție de frecvență este dată de expresia (VI.1,1).

Parametrii m și σ au o semnificație deosebită și anume m este media, iar σ^2 dispersia distribuției.

Intr-adevăr, efectuind aceeaşi substituţie ca mai sus, găsim

$$m=M(x)=\frac{1}{\sqrt{2\pi}\,\sigma}\int_{-\infty}^{+\infty}xe^{-\frac{(x-m)}{2\sigma^2}}dx=\frac{1}{\sqrt{\pi}}\int_{-\infty}^{+\infty}(m+\sqrt{2}\,\sigma y)\,e^{-y^2}dy=$$

$$=m+\frac{\sigma}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}d\,(e^{-y^2})=m+\frac{\sigma}{\sqrt{2\pi}}(e^{-y^2})_{-\infty}^{+\infty}=m,$$

$$\sigma^2=M(x-m)^2=\frac{1}{\sqrt{2\pi}\,\sigma}\int_{-\infty}^{+\infty}(x-m)^2e^{-\frac{(x-m)^2}{2\sigma^2}}dx=\frac{2\sigma^2}{\sqrt{\pi}}\int_{-\infty}^{+\infty}y^2e^{-y^2}dy=$$

$$=\frac{\sigma^2}{\sqrt{\pi}}\int_{-\infty}^{+\infty}yd\,(-e^{-y^2})=\frac{\sigma^2}{\sqrt{\pi}}\left[(-ye^{-y^2})_{-\infty}^{+\infty}+\int_{-\infty}^{+\infty}e^{-y^2}dy\right]=\sigma^2.$$

Laplace şi Gauss au arătat importanţa distribuţiei normale în studiul probabilistic a numeroase fenomene naturale. De aceea în literatura statistică ea este cunoscută sub numele de *distribuţia lui Laplace* sau *distribuţia lui Gauss*. Meritul priorităţii în această chestiune revine însă lui Laplace şi de aceea, considerăm că este mai îndreptăţit să se atribuie acestei distribuţii numele lui Laplace.

Reprezentarea grafică a funcţiei

$$f(x)=\frac{1}{\sqrt{2\pi}\,\sigma}e^{-\frac{(x-m)^2}{2\sigma^2}}$$

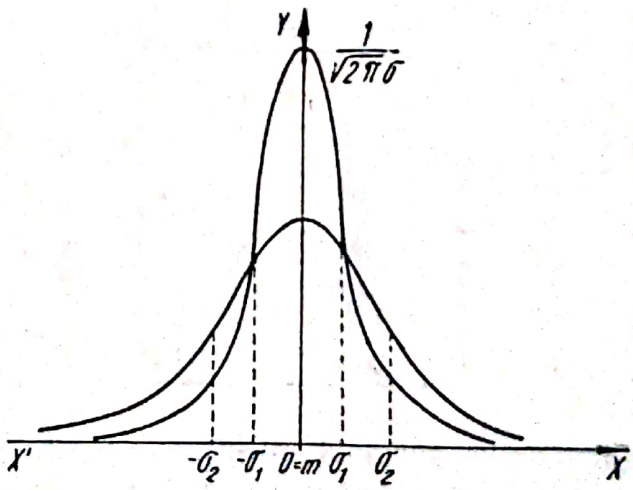


Fig. VI.1,1

ne arată că este simetrică faţă de axa y-lor. Pentru $x=m$, admite un maxim $f(m)=\frac{1}{\sqrt{2\pi}\cdot\sigma}$, iar în punctele de abscisă $x=\pm\sigma$, curba îşi schimbă semnul concavităţii, adică are două puncte de inflexiune.

Cu cît $|\sigma|$ este mai mic, cu atît $\max \frac{1}{\sqrt{2\pi}\sigma}$ şi

$$P(-A<x<A)=\frac{1}{\sqrt{2\pi}\sigma}\int_{-A}^{+A} e^{-\frac{x^2}{2}}\,dx$$

este mai mare.

Distribuţia normală fiind continuă se înţelege de la sine că nu este o distribuţie experimentală, directă, deoarece volumul oricărei distribuţii care rezultă din experienţe este finit.

Se poate constata, însă, că numeroase distribuţii care rezultă din experienţă se supun teoretic unei legi normale prin faptul că frecvenţele empirice sînt foarte apropiate de cele date de această lege.

Cunoscutul exemplu de mai jos, al lui Quetelet, ne arată acest lucru (tabela VI.1,1).

TABELA VI.1,1

Distribuţia în armata Potomakului în raport cu circumferinţa toracelui măsurat în ţoli

x	frecvenţe empirice	frecvenţe teoretice (prob.)	x	frecvenţe empirice	frecvenţe teoretice (prob.)
28	1	1	36	166	168
29	3	3	37	119	120
30	11	12	38	68	69
31	36	32	39	28	31
32	67	69	40	13	12
33	119	121	41	4	3
34	160	169	42	1	1
35	204	189	Total 1 000		1 000

Distribuţia normală prezintă o mare importanţă în statistica matematică, deoarece ea se întîlneşte foarte frecvent în aplicaţii.

Cercetarea legii limită a sumelor de variabile aleatorii

$$X_1+X_2+\ldots+X_n,$$

al căror studiu a fost început de Laplace şi care continuă şi astăzi să fie în centrul preocupărilor probabiliştilor, constituie o problemă care are o deosebită importanţă pentru studiul fenomenelor din natură prin metoda statistică.

Scanned with OKEN Scanner

Din această cauză, teorema care arată condiţiile în care suma de variabile aleatorii

$$X_1+X_2+\ldots+X_n$$

tinde către distribuţia normală poartă numele de *teorema centrală a calculului probabilităţilor.*

Cercetări mai moderne au arătat că independenţa variabilelor aleatorii care se însumează nu este o condiţie esenţială. Astfel, S. Bernştein a arătat că teorema lui Liapunov se menţine în condiţii destul de puţin restrictive şi pentru variabilele independente.

§ 2. Momente

Vom începe calculul momentelor distribuţiei normale $N(0, 1)$, adică ale distribuţiei normale cu media $m=0$ şi dispersia $\sigma^2=1$.

In acest caz, momentele obişnuite coincid cu momentele centrate. Avem, evident,

$$m_{2r+1}=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}x^{2r+1}e^{-\frac{x^2}{2}}dx=0,$$

deoarece funcţia $x^{2r+1}\cdot e^{-\frac{x^2}{2}}$ este o funcţie impară şi

$$m_{2r}=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}x^{2r}e^{-\frac{x^2}{2}}dx=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}x^{2r-1}d\left(-e^{-\frac{x^2}{2}}\right)=$$

$$=\frac{1}{\sqrt{2\pi}}\left[\left(-x^{2r-1}e^{-\frac{x^2}{2}}\right)_{-\infty}^{+\infty}+(2r-1)\int_{-\infty}^{+\infty}x^{2r-2}e^{-\frac{x^2}{2}}dx\right]=(2r-1)\,m_{2r-2}.$$

Dînd lui r valorile

$$r,\quad r-1,\ldots,2,\ 1,$$

rezultă următoarele relaţii de recurenţă

$$\begin{aligned}
m_{2r}&=(2r-1)\,m_{2r-2}\\
m_{2r-2}&=(2r-3)\,m_{2r-4}\\
m_{2r-4}&=(2r-5)\,m_{2r-6}\\
&\cdots\cdots\cdots\\
m_4&=3m_2\\
m_2&=m_0\qquad (m_0=1).
\end{aligned}$$

Inmulţind aceste relaţii, deducem

$$m_{2r}=1\cdot 3\ldots(2r-3)(2r-1)=\frac{(2r)!}{2^r\cdot r!}. \qquad \text{(VI.2,1)}$$

Momentele centrate ale distribuţiei $N(m,\sigma)$ se pot calcula imediat dacă se cunosc momentele distribuţiei normale $N(0,1)$.

Vom scrie

$$m_{2r+1}=\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{+\infty}(x-m)^{2r+1}e^{-\frac{(x-m)^2}{2\sigma^2}}dx=0,$$

deoarece funcţia $(x-m)^{2r+1}e^{-\frac{(x-m)^2}{2\sigma^2}}$ este, de asemenea, impară şi

$$m_{2r}=\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{+\infty}(x-m)^{2r}e^{-\frac{(x-m)^2}{2\sigma^2}}dx,$$

de unde făcînd substituţia $y=\frac{x-m}{\sigma}$ şi ţinînd seama de (VI.2,1) obţinem

$$m_{2r}=\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{+\infty}(x-m)^{2r}e^{-\frac{(x-m)^2}{2\sigma^2}}dx=\frac{\sigma^{2r}}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}y^{2r}e^{-\frac{y^2}{2}}dy=$$

$$=1\cdot 3\ldots(2r-1)\,\sigma^{2r}=\frac{(2r)!}{2^r\cdot r!}\sigma^{2r}. \qquad \text{(VI.2,2)}$$

Oblicitatea şi excesul pentru distribuţia normală $N(m,\sigma)$ rezultă imediat

$$\gamma_1=\frac{m_3}{(m_2)^{\frac{3}{2}}}=0,\quad \gamma_2=\frac{m_4}{m_2^2}-3=\frac{3\sigma^4}{\sigma^4}-3=0.$$

Prin urmare, oblicitatea şi excesul pentru legea normală cea mai generală sînt egale cu zero.

De aici, cunscutul procedeu al statisticii descriptive de a considera oblicitatea şi excesul ca nişte criterii pentru stabilirea normalităţii. Dacă aceste valori tipice sînt apropiate de zero, atunci este un indiciu că distribuţia diferă puţin de distribuţia normală.

După cum se observă, faptul că oblicitatea şi excesul sînt egale cu zero constituie o condiţie necesară, dar nu suficientă, şi numai în acest sens criteriul de mai sus trebuie aplicat în practică.

Scanned with OKEN Scanner

§ 3. Funcţia caracteristică

Ne vom ocupa, întocmai ca şi la calculul momentelor, în primul rînd, de determinarea funcţiei caracteristice a distribuţiei normale $N(0, 1)$ şi în al doilea rînd, de aceea a distribuţiei normale $N(m, \sigma)$. Astfel, funcţia caracteristică a distribuţiei normale $N(0, 1)$ este

$$\varphi(t)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} e^{xti}\,e^{-\frac{x^2}{2}}dx=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}\sum_{n=0}^{\infty}\frac{t^n i^n}{n!}x^n e^{-\frac{x^2}{2}}dx.$$

Seria de sub semnul integrală fiind uniform convergentă, vom putea să scriem mai departe

$$\varphi(t)=\sum_{n=0}^{\infty}\frac{t^n i^n}{n!}\cdot\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}x^n e^{-\frac{x^2}{2}}dx=\sum_{r=0}^{\infty}\frac{t^{2r}(-1)^r}{(2r)!}\cdot\frac{(2r)!}{2^r\cdot r!}=$$

$$=\sum_{r=0}^{\infty}(-1)^r\frac{\left(\frac{t^2}{2}\right)^r}{r!}=e^{-\frac{t^2}{2}}, \qquad \text{(VI.3,1)}$$

dacă $n=2r$ este un număr par şi avem în vedere relaţia (VI.2,2). Funcţia caracteristică a distribuţiei normale $N(m, \sigma)$ rezultă imediat, observînd că, în baza substituţiei $y=\frac{x-m}{\sigma}$, deducem

$$x=m+\sigma y.$$

Atunci, folosind expresia (VI.3,1) a funcţiei caracteristice a distribuţiei normale $N(0,1)$, vom avea

$$\psi(t)=M(e^{xti})=M\left[e^{(m+\sigma y)\,ti}\right]=e^{mti}M\left[e^{y(\sigma t)i}\right]=e^{mti}\varphi(\sigma t)=$$

$$=e^{mti-\frac{\sigma^2t^2}{2}}. \qquad \text{(VI.3,2)}$$

Teoremă. *Dacă variabilele aleatorii independente X_1 şi X_2 urmează fiecare, respectiv, legile normale $N(m_1, \sigma_1)$ şi $N(m_2, \sigma_2)$, atunci suma lor X_1+X_2 va urma legea normală $N(m, \sigma)$, unde $m=m_1+m_2$ şi $\sigma^2=\sigma_1^2+\sigma_2^2$.*

In adevăr, din teorema funcţiei caracteristice a sumei variabilelor independente rezultă

$$\psi(t)=e^{m_1 t i-\frac{\sigma_1^2 t^2}{2}}\, e^{m_2 t i-\frac{\sigma_2^2 t^2}{2}}=e^{(m_1+m_2)ti-\frac{(\sigma_1^2+\sigma_2^2)t^2}{2}},$$

care este funcţia caracteristică a legii normale $N(m,\ \sigma)$ cu $m=m_1+m_2$ şi $\sigma^2=\sigma_1^2+\sigma_2^2$, în virtutea relaţiei (VI.3,2).
Deci, teorema este demonstrată.
Această teoremă ne arată că distribuţia normală este o *distribuţie stabilă*.

§ 4. Polinoamele lui Hermite

In problema momentelor pe care am tratat-o (cap. IV.8,2), am arătat că, dacă avem o funcţie de frecvenţă $f(x)$, se pot construi nişte polinoame ortogonale, pentru care avem următoarele relaţii

$$\int_{-\infty}^{+\infty}\omega_m(x)\,\omega_n(x)f(x)\,dx=0 \qquad (m\neq n),$$
$$\int_{-\infty}^{+\infty}\omega_m^2(x)\,f(x)\,dx=1. \qquad \text{(VI.4,1)}$$

In cazul unei distribuţii oarecare, am stabilit că există o succesiune infinită de polinoame $\omega_n(x)$, $(n=1,\ 2,\ldots,\infty)$, care verifică relaţiile (VI.4,1).
Aceste polinoame intervin, după cum am văzut, la determinarea aproximativă a funcţiei de frecvenţă $f(x)$, atunci cînd sînt date primele m momente ale distribuţiei.
Pentru o distribuţie normală $N(0,\ 1)$ cu funcţia de frecvenţă

$$f(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}},$$

polinoamele $\omega_n(x)$ sînt *polinoamele lui Hermite*, definite prin egalitatea

$$\omega_n(x)=\frac{e^{x^2}}{(-2)^n}\cdot\frac{d^n e^{-x^2}}{dx^n}.$$

Din cauza legăturii lor cu distribuţia normală, polinoamele lui Hermite sînt des întrebuinţate în statistică.

Scanned with OKEN Scanner

§ 5. Distribuția normală multidimensională

Să considerăm vectorul

$$\begin{Vmatrix} x_1 \\ x_2 \\ \vdots \\ x_s \end{Vmatrix}$$

și matricea

$$A = \| a_{ij} \| \qquad (i, j = 1, 2, \ldots, s)$$

de rangul s. In acest caz determinantul $|A| \neq 0$.
Să considerăm, de asemenea, forma pătratică

$$x'Ax = \sum_{i=1}^{s} \sum_{j=1}^{s} a_{ij} x_i x_j, \tag{VI.5,1}$$

pe care o presupunem definită pozitivă.
Ne propunem să determinăm o constantă pozitivă h, astfel ca expresia

$$h \cdot e^{-x'Ax} \tag{VI.5,2}$$

să fie o funcție de frecvență într-un spațiu euclidian cu s dimensiuni. Pentru ca expresia (VI.5,2) să fie o funcție de frecvență este suficient ca să aibă loc egalitatea

$$h \int\int \ldots \int_{R_s} e^{-\sum_{i=1}^{s} \sum_{j=1} a_{ij} x_i x_j} dx_1 dx_2 \ldots dx_s = 1. \tag{VI.5,3}$$

Vom face o transformare ortogonală

$$x = Hy. \tag{VI.5,4}$$

Se știe că există întotdeauna o transformare ortogonală (VI.5,4), care reduce forma pătratică definită pozitivă (VI.5,1) la o sumă de s pătrate, adică

$$\sum_{i=1}^{s} \sum_{j=1}^{s} a_{ij} x_i x_j = \sum_{i=1}^{s} \lambda_i y_i^2,$$

unde $\lambda_i\ (i=1,\ 2,\ \ldots,\ s)$ sînt rădăcinile ecuației de gradul s

$$|A-\lambda E|=0,$$

E fiind matricea unitate.
In plus, valorile caracteristice λ_i trebuie să fie reale și pozitive, deoarece (VI.5,1) este o formă pătratică definită pozitivă.
Transformarea (VI.5,4), fiind ortogonală, determinantul funcțional corespunzător are valoarea 1 și, deci, egalitatea (VI.5,3) va deveni

$$h\iint\ldots\int_{R_s} e^{-\sum_{i=1}^{s}\lambda_i y_i^2}\,dy_1\,dy_2\ldots dy_s=1.$$

In această integrală multiplă, variabilele $y_i\ (i=1,\ 2,\ \ldots,\ s)$ se pot separa și, prin urmare, ea se mai poate scrie

$$h\prod_{i=1}^{s}\int_{-\infty}^{+\infty} e^{-\lambda_i y_i^2}\,dy_i=1: \qquad \text{(VI.5,5)}$$

Prin substituția

$$y_i=\frac{z_i}{\sqrt{\lambda_i}},$$

deducem din egalitatea (VI.5,5)

$$\frac{h}{\sqrt{\lambda_1\lambda_2\ldots\lambda_s}}\prod_{i=1}^{s}\int_{-\infty}^{+\infty} e^{-z_i^2}\,dz_i=\frac{h}{\sqrt{\lambda_1\lambda_2\ldots\lambda_s}}\cdot\pi^{\frac{s}{2}}=1,$$

de unde rezultă

$$h=\frac{\sqrt{\lambda_1\lambda_2\ldots\lambda_s}}{\pi^{\frac{s}{2}}}=\frac{\sqrt{|A|}}{\pi^{\frac{s}{2}}}, \qquad \text{(VI.5,6)}$$

ținînd seama că $\lambda_i\ (i=1,\ 2,\ \ldots,\ s)$ sînt rădăcinile ecuației

$$|A-\lambda E|=\begin{vmatrix} a_{11}-\lambda & a_{12} & \ldots & a_{1s} \\ a_{21} & a_{22}-\lambda & \ldots & a_{2s} \\ \ldots & \ldots & \ldots & \ldots \\ a_{s1} & a_{s2} & \ldots & a_{ss}-\lambda \end{vmatrix}=0.$$

Înlocuind în (VI.5,2) valoarea constantei h dată de relația (VI.5,6), obținem

$$\frac{\sqrt{|A|}}{\pi^{\frac{s}{2}}} e^{-x'Ax}. \qquad \text{(VI.5,7)}$$

Aceasta este *funcția de frecvență a distribuției normale cu s dimensiuni.*

Pentru a păstra analogia cu funcția de frecvență a distribuției normale cu o singură dimensiune $N(0, \sigma)$, care după cele stabilite mai înainte, este

$$\frac{1}{\sqrt{2\pi}\,\sigma} \cdot e^{-\frac{x^2}{2\sigma^2}},$$

vom considera funcția de frecvență (VI.5,7) a distribuției normale s-dimensionale în formă similara

$$\frac{\sqrt{|A|}}{(2\pi)^{\frac{s}{2}}} e^{-\frac{1}{2}x'Ax}. \qquad \text{(VI.5,8)}$$

De altfel, se putea ajunge în mod direct la forma (VI.5,8), prin efectuarea substituției $y_i = \frac{z_i}{\sqrt{2\lambda_i}}$ în loc de $y_i = \frac{z_i}{\sqrt{\lambda_i}}$.

Mai general, putem considera funcția de frecvență

$$\frac{\sqrt{|A|}}{(2\pi)^{\frac{s}{2}}} e^{-\frac{1}{2}(x'-m')A(x-m)}, \qquad \text{(VI.5,9)}$$

forma pătratică $(x'-m')A(x-m)$ fiind definită pozitivă, vectorul

$$m = \begin{Vmatrix} m_1 \\ m_2 \\ \vdots \\ m_s \end{Vmatrix}$$

şi m' transpusa lui, adică vectorul linie

$$m' = \| m_1, m_2, \ldots, m_s \| .$$

Vom vedea că vectorul m este valoarea medie a distribuţiei.
In cazul cînd matricea $A = \| a_{ij} \| \, (i, j = 1, 2, \ldots, s)$ are s linii şi s coloane, adică este de rangul s, vom spune că avem o *lege normală multidimensională propriu-zisă*, iar dacă matricea A este de rangul $r < s$, avem o *lege normală multidimensională singulară de rangul r*.

§ 6. Funcţia caracteristică a distribuţiei normale multidimensionale

Fie vectorul

$$t = \begin{Vmatrix} t_1 \\ t_2 \\ \vdots \\ t_s \end{Vmatrix}$$

şi transpusa lui

$$t' = \| t_1, t_2, \ldots, t_s \| .$$

Prin definiţie, vom numi funcţie caracteristică a distribuţiei normale multidimensionale, expresia

$$\varphi(t) = M(e^{it'x}),$$

sau, ţinînd seama de funcţia de frecvenţă (VI.5,8), vom scrie

$$\varphi(t) = \frac{\sqrt{|A|}}{(2\pi)^{\frac{s}{2}}} \int\int \ldots \int_{R_s} e^{it'x - \frac{1}{2} x'Ax} \, dx_1 dx_2 \ldots dx_s . \qquad \text{(VI.6,1)}$$

Să aplicăm din nou transformarea ortogonală

$$x = Hy,$$

care reduce forma pătratică $x'Ax$ la suma de pătrate $\sum_{j=1}^{s} \lambda_j y_j^2$.

In acelaşi timp, se ştie că, printr-o transformare oarecare $x = Hy$, vectorul t se transformă într-un vector u, astfel ca

$$t = (H')^{-1} \cdot u.$$

Transformarea fiind ortogonală, există relația

$$H' = H^{-1}$$

și deci

$$(H')^{-1} = (H^{-1})^{-1} = H,$$

de unde rezultă că

$$t = H \cdot u. \qquad \text{(VI.6,2)}$$

In baza egalității (VI.6,2), produsul $t'x$ devine

$$t'x = u'H' \cdot x = u'H' \cdot Hy = u'y.$$

Determinantul funcțional al transformării ortogonale este, evident, egal cu unu.

Atunci, în urma transformării de mai sus, funcția caracteristică (VI.6,1) va fi

$$\varphi(t) = \frac{\sqrt{|A|}}{(2\pi)^{\frac{s}{2}}} \int\int \cdots \int_{R_s} e^{iu'y - \frac{1}{2}\sum_{j=1}^{s} \lambda_j y_j^2} dy_1 dy_2 \ldots dy_s. \qquad \text{(VI.6,3)}$$

Produsul matriceal $u'y$ fiind egal cu $u_1y_1 + u_2y_2 + \ldots + u_sy_s$, vom putea separa variabilele în integrala multiplă (VI.6,3), care, astfel, se va scrie

$$\varphi(t) = \frac{\sqrt{|A|}}{(2\pi)^{\frac{s}{2}}} \prod_{j=1}^{s} \int_{-\infty}^{+\infty} e^{iu_j'y_j - \frac{1}{2}\lambda_j y_j^2} dy_j =$$

$$= \frac{\sqrt{|A|}}{(2\pi)^{\frac{s}{2}}} \cdot \frac{(2\pi)^{\frac{s}{2}}}{\sqrt{\lambda_1\lambda_2 \ldots \lambda_s}} e^{-\frac{1}{2}\sum \frac{u_j^2}{\lambda_j}} = e^{-\frac{1}{2}\sum_{j=1}^{s} \frac{u_j^2}{\lambda_j}}. \qquad \text{(VI.6,4)}$$

Să exprimăm funcția caracteristică (VI.6,4) în funcție de vectorul t și transpusa lui, t'.

Vom pleca de la egalitatea care are loc pentru orice transformare ortogonală $x = Hy$

$$\sum_{j=1}^{s} \lambda_j y_j^2 = x'Ax = y'H'A\,Hy = y'By,$$

unde avem

$$B=H'AH=\begin{Vmatrix}\lambda_1 & 0 & 0\ldots 0\\ 0 & \lambda_2 & 0\ldots 0\\ \ldots & \ldots & \ldots\\ 0 & 0 & 0 \quad \lambda_s\end{Vmatrix}.$$

In acest caz, vom avea, de asemenea, egalitatea

$$\sum_{j=1}^{s}\lambda_j u_j^2=u'Bu$$

şi de aici

$$\sum_{j=1}^{s}\frac{u_j^2}{\lambda_j}=u'B^{-1}u, \qquad \text{(VI.6,5)}$$

deoarece

$$B^{-1}=\begin{Vmatrix}\frac{1}{\lambda_1} & 0 & 0\ldots 0\\ 0 & \frac{1}{\lambda_2} & 0\ldots 0\\ \ldots & \ldots & \ldots\\ 0 & 0 & 0\ldots\frac{1}{\lambda_s}\end{Vmatrix}.$$

De la forma pătratică definită pozitivă $u'B^{-1}u$, trecem la o formă pătratică definită pozitivă de forma $t'Ct$, unde matricea C trebuie determinată.

In adevăr, din relaţiile stabilite

$$t=Hu,$$

$$B=H'AH,$$

deducem relaţiile următoare

$$H^{-1}t=H^{-1}Hu=u,$$

$$t'=u'H'=u'H^{-1},$$

$$t'H=u'H^{-1}H=u',$$

$$B^{-1}=(H'AH)^{-1}=(AH)^{-1}H=H^{-1}\cdot A^{-1}\cdot H.$$

Scanned with OKEN Scanner

Inlocuind în (VI.6,5) matricele u', B^{-1} şi u prin noile lor expresii, obţinem

$$\sum_{j=1}^{s} \frac{u_j^2}{\lambda_j} = t'HH^{-1}A^{-1}HH^{-1}t = t'A^{-1}t.$$

Ţinînd seama de acest rezultat, funcţia caracteristică (VI.6,4) a distribuţiei normale multidimensionale este

$$\varphi(t) = e^{-\frac{1}{2}t'A^{-1}t}. \qquad \text{(VI.6,6)}$$

După cum vom vedea, matricea A^{-1}, care intră în alcătuirea funcţiei caracteristice, coincide cu matricea covariantă a distribuţiei. De aceea, se va numi matrice covariantă a distribuţiei.

Cunoscînd acum funcţia caracteristică a distribuţiei normale multidimensionale cu funcţia de frecvenţă

$$\frac{\sqrt{|A|}}{(2\pi)^{\frac{s}{2}}} \cdot e^{-x'Ax},$$

vom putea să determinăm funcţia caracteristică a distribuţiei normale multidimensionale

$$\frac{\sqrt{|A|}}{(2\pi)^{\frac{s}{2}}} e^{(x'-m')A(x-m)}.$$

Considerînd transformarea

$$x = y + m,$$

funcţia caracteristică a vectorului x, potrivit definiţiei sale, este

$$\psi(t) = M(e^{it'x}) = M(e^{it'(y+m)}) = e^{it'm} M(e^{it'y}) = e^{it'm - \frac{1}{2}t'A^{-1}t}, \qquad \text{(VI.6,7)}$$

ştiind că $M(e^{it'y})$ este funcţia caracteristică a cărei expresie este dată de (VI.6,6).

Distribuţia normală multidimensională este o *distribuţie stabilă*, ceea ce rezultă din teorema de mai jos.

Teoremă. *Dacă variabilele independente* ξ *şi* η *urmează fiecare o lege normală cu s dimensiuni propriu-zisă, avînd respectiv va-*

lorile medii a şi b şi matricele covariante A^{-1} *şi* B^{-1}, *atunci variabila* $\xi+\eta$ *va urma o lege normală cu s dimensiuni propriu-zisă, cu valoarea medie* $a+b$ *şi matricea covariantă* $(A+B)^{-1}$.

Demonstraţia este imediată, folosind teorema funcţiei caracteristice a sumei de variabile independente

$$\psi_1(t)\psi_2(t)=e^{it'a-\frac{1}{2}t'A^{-1}t}\cdot e^{it'b-\frac{1}{2}t'B^{-1}t}=e^{it'(a+b)-\frac{1}{2}t'(A+B)^{-1}t}.$$

Aceasta este, însă, funcţia caracteristică a distribuţiei normale multidimensionale cu valoarea medie $a+b$ şi matricea covariantă $(A+B)^{-1}$ şi, prin urmare, teorema este demonstrată.

Vom demonstra încă o teoremă, care scoate în evidenţă o altă proprietate esenţială a legii normale multidimensionale. Teorema ne va arăta că o lege normală multidimensională se păstrează printr-o transformare liniară $y=Tx$.

Teoremă. *Dacă vectorul aleator* $x(x_1, x_2 \ldots, x_s)$ *urmează legea normală s — dimensională, cu valoarea medie m şi matricea covariantă* A^{-1}, *atunci vectorul aleator* $y(y_1, y_2, \ldots, y_s)$, *obţinut prin transformarea* $y=Tx$, $|T|\neq 0$, *urmează o lege normală s-dimensională, cu valoarea medie Tm şi matricea covariantă* $TA^{-1}T'$.

Pentru demonstraţie, vom folosi, de asemenea, funcţia caracteristică, a vectorului aleator $y(y_1, y_2, \ldots, y_s)$, care se scrie

$$\psi(t)=M(e^{it'y})=M(e^{it'Tx})=M(e^{i(T't)'x}),$$

folosind proprietatea operaţiei de transpunere $(T't)'=t'(T')'=t'T$. Notînd $u=T't$, funcţia caracteristică $\psi(t)$ devine

$$\psi(t)=M(e^{iu'x})=e^{iu'm-\frac{1}{2}u'A^{-1}u}.$$

Revenind la vectorul t, obţinem

$$\psi(t)=e^{i(T't)'m-\frac{1}{2}(T't)'A^{-1}(T't)}=e^{it'Tm-\frac{1}{2}t'TA^{-1}T't}.$$

După cum se observă, funcţia caracteristică găsită este funcţia caracteristică a unei distribuţii normale s — dimensionale cu valoarea medie Tm şi matricea covariantă $(TA^{-1}T')^{-1}$, ceea ce demonstrează teorema.

Scanned with OKEN Scanner

§ 7. Momentele distribuției normale multidimensionale

Momentele distribuției normale multidimensionale se calculează cu ajutorul funcției caracteristice.

Astfel, ne propunem să determinăm momentele distribuției normale cu s dimensiuni a cărei funcție de frecvență este

$$\frac{\sqrt{|A|}}{(2\pi)^{\frac{s}{2}}} \cdot e^{-\frac{1}{2}x'Ax}.$$

Funcția caracteristică corespunzătoare acestei distribuții este

$$\varphi(t) = e^{-\frac{1}{2}t'A^{-1}t} = e^{-\frac{1}{2}\sum_{h=1}^{s}\sum_{k=1}^{s}a_{hk}^{-1}t_h t_k}.$$

Din aceasta deducem

$$M_h = \frac{1}{i}\left(\frac{\partial\varphi}{\partial t_h}\right)_{t_1, t_2, \ldots, t_s=0} =$$

$$= \left[ie^{-\frac{1}{2}\sum_{h=1}^{s}\sum_{k=1}^{s}a_{hk}^{-1}t_h t_k}\sum_{k=1}^{s}a_{hk}^{-1}t_k\right]_{t_1, t_2, \ldots, t_s=0} = 0,$$

$$M_{hk} = \frac{1}{i^2}\left(\frac{\partial^2\varphi}{\partial t_h \partial t_k}\right)_{t_1, t_2, \ldots, t_s=0} =$$

$$= \left[i^2 e^{-\frac{1}{2}\sum_{h=1}^{s}\sum_{k=1}^{s}a_{hk}^{-1}t_h t_k}\sum_{k=1}^{s}a_{hk}^{-1}\cdot t_k \cdot \sum_{h=1}^{s}a_{hk}^{-1}t_h + \right. \tag{VI.7,1}$$

$$\left. + a_{hk}^{-1}e^{-\frac{1}{2}\sum_{h=1}^{s}\sum_{k=1}^{s}a_{hk}^{-1}t_h t_k}\right]_{t_1, t_2, \ldots, t_s=0} = a_{hk}^{-1} = \frac{A_{hk}}{|A|},$$

. . . . . .

Momentele distribuţiei normale cu s dimensiuni a cărei funcţie de frecvenţă este

$$\frac{\sqrt{|A|}}{(2\pi)^{\frac{s}{2}}} e^{-\frac{1}{2}(x'-m')A(x-m)}$$

se vor calcula după aceeaşi metodă.
Deoarece avem pentru funcţia caracteristică expresia

$$\psi(t) = e^{it'm-\frac{1}{2}t'A^{-1}t} = e^{i\sum_{h=1}^{s} t_h m_h - \frac{1}{2}\sum_{h=1}^{s}\sum_{k=1}^{s} a_{hk}^{-1} t_h t_k},$$

obţinem

$$M_h = \frac{1}{i}\left(\frac{\partial\psi}{\partial t_h}\right)_{t_1, t_2, \ldots, t_s=0} =$$

$$= \left[\left(m_h + i\sum_{k=1}^{s} a_{hk}^{-1} t_k\right) e^{i\sum_{h=1}^{s} t_h m_h - \frac{1}{2}\sum_{h=1}^{s}\sum_{k=1}^{s} a_{hk}^{-1} t_h t_k}\right]_{t_1, t_2, \ldots, t_s=0} = m_h,$$

$$M_{hk} = \frac{1}{i^2}\left(\frac{\partial^2\psi}{\partial t_h \partial t_k}\right)_{t_1, t_2, \ldots, t_s=0} =$$

$$= \left[\left(m_h + i\sum_{k=1}^{s} a_{hk}^{-1} t_k\right)\left(m_k + i\sum_{h=1}^{s} a_{hk}^{-1} t_k\right) \times \right.$$

$$\times e^{i\sum_{h=1}^{s} t_h m_h - \frac{1}{2}\sum_{h=1}^{s}\sum_{k=1}^{s} a_{hk}^{-1} t_h t_k} +$$

$$\left. + a_{hk}^{-1} e^{i\sum_{h=1}^{s} t_h m_h - \frac{1}{2}\sum_{h=1}^{s}\sum_{k=1}^{s} a_{hk}^{-1} t_h t_k}\right]_{t_1, t_2, \ldots, t_s=0} = m_h m_k + a_{hk}^{-1}. \quad \text{(VI.7,2)}$$

. . . . . .

In particular, dacă $m_j = 0$ $(j = 1, 2, \ldots, s)$, momentele (VI.7,2) coincid cu momentele (VI.7,1).

In același timp observăm că vectorul m ale cărui componente sînt $m_1, m_2, \ldots, m_s$ reprezintă valoarea medie a distribuției normale s — dimensionale cu funcția de frecvență

$$\frac{\sqrt{|A|}}{(2\pi)^{\frac{s}{2}}} e^{-\frac{1}{2}(x-m)' A (x-m)}.$$

Totodată, vom avea

$$M(x_h - m_h) = 0,$$

$$M(x_h - m_h)^2 = m_h^2 + a_{hh}^{-1}, \qquad \text{(VI.7,3)}$$

$$M(x_h - m_h)(x_k - m_k) = M_{hk} - m_h m_k = a_{hk}^{-1} = \frac{A_{hk}}{|A|},$$

. . . . . .

Acum se explică justificarea denumirii de matrice covariantă a matricei $A^{-1} = \| a_{hk}^{-1} \|$ $(h, k = 1, 2, \ldots, s)$.

Mai observăm că momentele M_{hk}, $M(x_h - m_h)(x_k - m_k), \ldots$ ale distribuției normale cu s dimensiuni, date de expresiile (VI.7,1), (VI.7,2), (VI.7,3), există numai dacă $|A| \neq 0$, adică în cazul distribuției multidimensionale propriu-zisă.

In cazul particular al distribuției normale cu două dimensiuni cu centrul în (m_1, m_2) și valorile medii

$$m_{20} = \sigma_1^2, \quad m_{11} = \rho\sigma_1\sigma_2, \quad m_{02} = \sigma_2^2,$$

rezultă că matricea covariantă A^{-1} este

$$A^{-1} = \begin{Vmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{Vmatrix}.$$

Prin urmare

$$A = \begin{Vmatrix} \frac{1}{(1-\rho^2)\sigma_1^2} & -\frac{\rho}{(1-\rho^2)\sigma_1\sigma_2} \\ -\frac{\rho}{(1-\rho^2)\sigma_1\sigma_2} & \frac{1}{(1-\rho^2)\sigma_2^2} \end{Vmatrix},$$

$$|A| = \frac{1}{(1-\rho^2)\sigma_1^2\sigma_2^2},$$

Prin urmare, funcţia de frecvenţă a distribuţiei normale cu două dimensiuni, în baza formulei generale, este

$$f(x,\ y)=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\cdot e^{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-m_1)^2}{\sigma_1^2}-\frac{2\rho\,(x-m_1)\,(y-m_2)}{\sigma_1\sigma_2}+\frac{(y-m_2)^2}{\sigma_2^2}\right]}.$$

Funcţia caracteristică a acestei distribuţii este

$$\varphi(u,\ v)=e^{i\,(m_1u+m_2v)-\frac{1}{2}\,(\sigma_1^2u^2+2\rho\sigma_1\sigma_2uv+\sigma_2^2v^2)}.$$

Capitolul VII

Distribuţia binomială

§ 1. Definiţii

Să considerăm un cîmp de probabilitate finit, care constă în evenimentele A şi non A (evenimentul contrar A'), în afară de evenimentul total E şi evenimentul imposibil Φ. Experienţa care, de fiecare dată, duce la apariţia evenimentului A sau a evenimentului non A, o vom numi *probă*.

n probe independente dau naştere fiecare la unul din cele două evenimente A şi non A, cu probabilităţile constante p şi q în fiecare probă. Intre probabilităţile p şi q există, evident, relaţia

$$p+q=1,$$

deoarece $A+A'=E$ şi $A\cdot A'=\Phi$.

Presupunem că în cele n probe independente, evenimentul A s-a verificat de α ori, iar evenimentul contrar A' de $(n-\alpha)$ ori.

Distribuţia variabilei aleatorii α se numeşte *distribuţie binomială* din cauza legăturii formale pe care o are, după cum vom vedea mai departe, cu binomul lui Newton.

In aplicaţiile statisticii matematice, întîlnim deseori asemenea distribuţii ale căror variabile iau o succesiune de valori corespunzătoare repetării unui eveniment de oricîte ori, în condiţii identice, în mod independent.

Scanned with OKEN Scanner

Exemplul 1. Dacă aruncăm o monedă de 100 ori, fiecare aruncare este o probă care că naştere la apariţia uneia dintre cele două feţe ale monedei. Numărul apariţiilor uneia dintre feţe în cele 100 aruncări este o variabilă aleatorie, care poate lua valorile 0, 1, 2, ..., 100.

Probele sînt independente şi în fiecare probă variabila ia numai valorile 1 sau 0, după cum s-a verificat sau nu apariţia uneia dintre cele două feţe cu probabilităţile egale cu jumătate fiecare.

Pentru fiecare probă avem distribuţia

$$X_i \begin{pmatrix} 1 & 0 \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix} \qquad (i=1, 2, \ldots, 100).$$

Exemplul 2. O serie de n piese obţinute de la aceeaşi maşină, pentru care avem probabilitatea constantă p ca fiecare piesă să fie bună, din punct de vedere tehnic, poate fi considerată ca o succesiune de probe independente, de tipul de mai sus, cu condiţia ca fabricarea unei piese să nu influenţeze asupra fabricării alteia.

In cele n probe, numărul pieselor, care corespund din punct de vedere tehnic, este o variabilă aleatorie, care poate lua valorile $0, 1, 2, \ldots, n$. Intr-o singură probă, variabila ia valorile 1 sau 0, după cum piesa este executată bine sau nu, cu aceleaşi probabilităţi p sau $1-p$.

Există, deci, următoarea distribuţie pentru fiecare probă.

$$X_i \begin{pmatrix} 1 & 0 \\ p & 1-p \end{pmatrix}.$$

Exemplul 3. Să presupunem, acum, că există probabilităţile p_x şi q_x ca o persoană în vîrstă de x ani să fie în viaţă după trecerea unui an, sau să moară în cursul acestui an.

Considerînd o populaţie de N persoane, avînd fiecare aceeaşi vîrstă x, care au condiţii de viaţă suficient de apropiate ca să admitem aceleaşi probabilităţi de viaţă sau de moarte p_x şi q_x, ajungem iarăşi la o serie de probe independente de tipul arătat.

Intr-adevăr, cercetînd mortalitatea acestui grup de N persoane, numărul persoanelor în viaţă după un an este o variabilă aleatorie care poate lua valorile $0, 1, 2, \ldots, N$. O singură persoană poate să fie în viaţă sau să moară în timpul anului, ceea ce înseamnă că variabila aleatorie în cazul unei singure probe ia valorile 1 sau zero cu probabilităţile constante p_x şi q_x.

Rezultă, prin urmare, ca distribuţie pentru probele individuale

$$X_i \begin{pmatrix} 1 & 0 \\ p_x & q_x \end{pmatrix} \qquad (i=1, 2, \ldots, N).$$

Distribuţia binomială este una dintre cele mai vechi din calculul probabilităţilor.

J. Bernoulli a studiat temeinic distribuţia binomială, cu ajutorul ei a găsit şi demonstrat legea numerelor mari, de care i se leagă numele şi a deschis

Scanned with OKEN Scanner

drumul aplicațiilor calculului probabilităților în diverse ramuri ale științei. De aceea, pe drept cuvînt, J. Bernoulli (1654-1705) poate fi considerat ca cel dintîi mare probabilist al lumii.

Schema pe care a dat-o și care servește la studiul multor fenomene statistice este realizată de o succesiune de extracții independente dintr-o urnă, care conține a bile albe și n bile negre, a și n fiind constante în decursul extracțiilor succesive.

Probabilitatea de a scoate o bilă albă sau neagră este aceeași la fiecare probă și anume

$$p=\frac{a}{a+n}, \quad q=\frac{n}{a+n}.$$

Schema imaginată de Bernoulli este foarte sugestivă și a jucat un rol important în dezvoltarea statisticii matematice.

§ 2. Probabilitatea $P_{n,\alpha}$

Teoremă. Se fac n probe independente astfel că, în fiecare probă, probabilitatea unui eveniment dat este p, iar a evenimentului contrar este q. *Probabilitatea ca în cele n probe, primul eveniment să se verifice de α ori, deci al doilea eveniment de $(n-\alpha)$ ori este*

$$P_{n,\alpha}=C_n^{\alpha} p^{\alpha} q^{n-\alpha}.$$

Notăm cu A primul eveniment și cu A' evenimentul contrar.

Numărul de probe în care se verifică evenimentul A este o variabilă aleatorie x care poate lua valorile frecvenței evenimentului A

$$0, 1, 2, \ldots, \alpha, \ldots, n$$

cu probabilitățile respective

$$P_{n,0}, P_{n,1}, P_{n,2}, \ldots, P_{n,\alpha}, \ldots, P_{n,n}.$$

Ne propunem, deci, să determinăm probabilitatea $P_{n,\alpha}$ de a se verifica evenimentul A de α ori, iar evenimentul A' de $(n-\alpha)$ ori în cele n probe independente.

Dacă cele două evenimente A și A' se vor succede în ordinea următoare

$$\underbrace{AAA\ldots A}_{\alpha \text{ ori}} \quad \underbrace{A'A'A'\ldots A'}_{(n-\alpha) \text{ ori}},$$

atunci, în baza teoremei probabilităţilor compuse, probabilitatea evenimentului ca A şi A' să se verifice în ordinea de mai sus este

$$p^{\alpha} q^{n-\alpha}.$$

Cum însă ordinea de verificare a evenimentelor A şi A' poate fi oarecare, putem, în baza teoremei probabilităţilor totale, să descompunem evenimentul de a se verifica A de α ori şi A' de $(n-\alpha)$ ori în orice ordine, în evenimente de probabilităţi egale cu $p^{\alpha} q^{n-\alpha}$ fiecare.

Numărul acestor evenimente va fi, evident,

$$C_n^{\alpha} = \frac{n!}{\alpha!\,(n-\alpha)!}.$$

Rezultă, astfel, pentru probabilitatea căutată

$$P_{n,\alpha} = C_n^{\alpha} p^{\alpha} q^{n-\alpha}. \qquad \text{(VII.2,1)}$$

Observăm că expresia (VII.2,1) a probabilităţii $P_{n,\alpha}$ este termenul general al dezvoltării binomului lui Newton, de unde şi justificarea numelui de *distribuţie binomială*

$$(p+q)^n = p^n + C_n^1 p^{n-1} q + \ldots + C_n^i p^{n-i} q^i + \ldots + q^n. \qquad \text{(VII.2,2)}$$

Mai observăm că, suma primilor α termeni în dezvoltarea (VII.2,2) a binomului lui Newton reprezintă probabilitatea ca evenimentul A să se verifice de cel puţin $(n-\alpha+1)$ ori, iar evenimentul A' de cel mult $(\alpha-1)$ ori în n probe.

Distribuţia variabilei aleatorii x, deci a frecvenţei evenimentului A în cele n probe independente, este indicată în tabloul următor

$$X\begin{pmatrix} 0, & 1, & 2 & , \ldots , & \alpha & , \ldots , & n \\ q^n, & C_n^1 p q^{n-1}, & C_n^2 p^2 q^{n-2}, \ldots , & C_n^{\alpha} p^{\alpha} q^{n-\alpha}, \ldots , & & & C_n^n p^n \end{pmatrix}.$$

Mai putem da, pentru determinarea probabilităţii $P_{n,\alpha}$, demonstraţia următoare.

Să descompunem evenimentul A care se verifică de α ori şi A' de $(n-\alpha)$ ori în n probe independente, în alte două evenimente.

1. Evenimentul care constă în faptul că în prima probă se verifică A, iar în restul de $(n-1)$ probe se verifică A' de $(\alpha-1)$ ori.

Scanned with OKEN Scanner

2. Evenimentul care constă în faptul că în prima probă se verifică A', iar în restul de $(n-1)$ probe se verifică A de α ori.
Atunci, în baza teoremelor probabilităților compuse și totale, avem relația de recurență

$$P_{n,\alpha}=pP_{n-1,\alpha-1}+qP_{n-1,\alpha}. \qquad \text{(VII.2,3)}$$

Observînd că în relația (VII.2,3) n poate lua oricare dintre valorile $1, 2, \ldots, n$ iar α oricare dintre valorile $0, 1, 2, \ldots, n$, obținem, aplicînd din nou teorema probabilităților totale,

$$\sum_{i=0}^{1} P_{1,i}=p+q,$$

$$\sum_{i=0}^{2} P_{2,i}=p^2+C_2^1pq+q^2=(p+q)^2,$$

$$\sum_{i=0}^{3} P_{3,i}=p^3+C_3^1p^2q+C_3^2pq^2+q^3=(p+q)^3, \qquad \text{(VII.2,4)}$$

$$\sum_{i=0}^{4} P_{4,i}=p^4+C_4^1p^3q+C_4^2p^2q^2+C_4^3pq^3+q^4=(p+q)^4,$$

. . . . . .

Pentru a arăta că egalitățile (VII.2,4) sînt valabile, oricare ar fi n întreg, vom presupune că ele sînt valabile pentru $n-1$.

Ținînd seama de relația (VII.2,3) și că $\sum_{i=0}^{n-1} P_{n-1,i}=(p+q)^{n-1}$, putem scrie

$$\sum_{i=0}^{n} P_{n,i}=\sum_{i=0}^{n}(pP_{n-1,i-1}+qP_{n-1,i})=p(p+q)^{n-1}+q(p+q)^{n-1}=$$

$$=(p+q)^n=\sum_{i=0}^{n} C_n^i p^i q^{n-i}, \qquad \text{(VII.2,5)}$$

unde $P_{n-1,-1}$ și $P_{n-1,n}$ sînt, evident, egale cu zero.
Deoarece, egalitățile (VII.2,4) sînt verificate pentru primele valori întregi ale lui n, rezultă că sînt adevărate pentru orice n întreg.
Din egalitățile (VII.2,5) deducem imediat

$$P_{n,\alpha}=C_n^\alpha p^\alpha q^{n-\alpha}.$$

Scanned with OKEN Scanner

§ 3. Variația probabilităților $P_{n,\alpha}$

Fie dat șirul de probabilități

$$P_0,\ P_1,\ P_2,\ \ldots,\ P_n, \tag{VII.3,1}$$

unde am notat, pentru prescurtare, cu P_i probabilitatea $P_{n,i}$ $(i=0,\ 1,\ \ldots,\ n)$.

În șirul considerat, cea mai mare dintre probabilitățile P_i corespunde unei valori λ, care în cele n probe va apare în mod teoretic de cele mai multe ori față de celelalte valori $i \neq \lambda$ ale șirului.

Valoarea λ se numește *valoarea cea mai probabilă.*

Teoremă. *Valoarea cea mai probabilă* λ *satisface inegalitățile*

$$np-q<\lambda<np+p.$$

Am văzut că probabilitatea P_α este dată de expresia

$$C_n^\alpha p^\alpha q^{n-\alpha} \quad (\alpha=0,\ 1,\ \ldots,\ n). \tag{VII.3,2}$$

Presupunînd că expresiile (VII.3,2) sînt aranjate după puterile descrescătoare ale lui p, atunci raportul uneia dintre ele față de precedenta va fi

$$\frac{C_n^\alpha p^\alpha q^{n-\alpha}}{C_n^{\alpha+1} p^{\alpha+1} p^{n-\alpha-1}} = \frac{\alpha+1}{n-\alpha} \cdot \frac{q}{p}. \tag{VII.3,3}$$

Înlocuind în relația (VII.3,3) pe α cu $\alpha-1$, obținem raportul

$$\frac{\alpha}{n-\alpha+1} \cdot \frac{q}{p}. \tag{VII.3,4}$$

Pentru determinarea valorii lui λ trebuie să aibă loc inegalitățile

$$\frac{\alpha+1}{n-\alpha} \cdot \frac{q}{p} > 1, \quad \frac{\alpha}{n-\alpha+1} \cdot \frac{q}{p} < 1$$

sau

$$np-q<\alpha<np+p. \tag{VII.3,5}$$

Numărul întreg α, care satisface inegalitățile (VII.3,5), este valoarea cea mai probabilă λ. Această valoare este, astfel, în mod unic determinată.

Numărul $np+p$ nu este, în general, un număr întreg. În cazul acesta, nici numărul $np-q$ nu este un număr întreg, deoarece diferă de $np+p$ prin $p+q=1$.

In cazul excepţional, cînd $np+p$ este un număr întreg, vom avea pentru λ două valori. Pentru că $np-q$ este de asemenea un număr întreg, cele două valori ale lui λ vor corespunde probabilităţilor maxime

$$P_{np-q}=P_{np+p}.$$

Din inegalităţile (VII.3,5) rezultă că valoarea cea mai probabilă este foarte apropiată de np, care, după cum vom vedea, este tocmai valoarea medie a lui α. Prin urmare, pentru distribuţia binomială, *modul* diferă foarte puţin de *valoarea medie*.

Variaţia probabilităţilor P_α capătă o anumită regularitate o dată cu creşterea lui n. Tabelele VII.3,1 şi VII.3,2 scot în evidenţă acest lucru.

Tabela VII.3,1 ne arată că distribuţia binomială pentru 20 probe este asimetrică, cu excepţia cazului $p=q=0{,}5$ şi că această asime-

TABELA VII.3,1

Valorile probabilităţilor P_α, calculate pe baza valorilor diferite ale probabilităţilor p şi q într-un număr de 20 probe, potrivit schemei lui J. Bernoulli. Valorile probabilităţilor P_α sînt înmulţite cu 10 000

α	$p=0{,}1$ $q=0{,}9$	$p=0{,}2$ $q=0{,}8$	$p=0{,}3$ $q=0{,}7$	$p=0{,}4$ $q=0{,}6$	$p=0{,}5$ $q=0{,}5$
0	1 216	115	8	—	—
1	2 702	576	68	5	—
2	2 852	1 369	268	31	2
3	1 901	2 054	716	123	11
4	898	2 182	1 304	350	46
5	319	1 746	1 789	746	148
6	89	1 091	1 916	1 244	370
7	20	545	1 643	1 659	739
8	4	222	1 144	1 797	1 201
9	1	74	654	1 597	1 602
10		20	308	1 171	1 762
11		5	120	710	1 602
12		1	39	355	1 201
13			10	146	739
14			2	49	370
15				13	148
16				3	46
17					11
18					2
19					—
20					—

Scanned with OKEN Scanner

TABELA VII.3,2

Valorile probabilităţilor 10 000 P_a calculate pe baza probabilităţilor $p=0,1$ şi $q=0,9$ pentru un număr diferit de probe, conform schemei lui J. Bernoulli

a	$n=5$	$n=10$	$n=20$	$n=50$	$n=100$
0	5 905	3 487	1 216	51	—
1	3 281	3 874	2 702	286	3
2	729	1 937	2 852	779	16
3	81	574	1 901	1 386	59
4	4	112	898	1 809	159
5	—	14	319	1 849	338
6	—	2	89	1 541	596
7	—	—	20	1 076	889
8	—	—	4	643	1 148
9	—	—	1	333	1 304
10			—	155	1 319
11			—	61	1 199
12			—	22	988
13			—	7	743
14				2	513
15				—	327
16				—	193
17				—	106
18				—	54
19				—	26
20					12
21					5
22					2
23					1

trie este cu atît mai pronunţată cu cît p este mai mic. Totuşi, dacă n creşte, asimetria devine din ce în ce mai puţin pronunţată. Aceasta rezultă din tabela VII.3,2, din care rezultă că, deşi probabilitatea p este egală cu 0,1, totuşi pentru 100 probe se ajunge la o distribuţie simetrică.

§ 4. Funcţia de distribuţie

Funcţia de distribuţie a distribuţiei binomiale este funcţia $F(x)$ definită de relaţia

$$F(x)=P(\alpha \leqq x)=\sum_{a=0}^{[x]} P_{n,a}, \qquad \text{(VII.4,1)}$$

unde am notat cu $[x]$ cel mai mare întreg conţinut în x.

Scanned with OKEN Scanner

Potrivit acestei definiţii, funcţia de repartiţie $F(x)$ a distribuţiei binomiale corespunde probabilităţilor cumulate $\sum_{\alpha=0}^{[x]} P_{n,\alpha}$ şi este o funcţie cu trepte, admiţînd discontinuităţi de prima speţă în fiecare dintre punctele 0, 1, 2, ..., $[x]$.

Avem astfel valorile de mai jos ale funcţiei de repartiţie a distribuţiei binomiale, pentru variabila aleatorie α, care poate lua valorile 0, 1, 2, ..., n:

$$F_n(x)=\begin{cases} =0, & x<0, \\ =C_n^0 q^n, & 0\leqq x<1, \\ =\sum\limits_{i=0}^{1} C_n^i p^i q^{n-i}, & 1\leqq x<2, \\ =\sum\limits_{i=0}^{2} C_n^i p^i q^{n-i}, & 2\leqq x<3, \\ \cdots\cdots\cdots & \\ \sum\limits_{i=0}^{n-1} C_n^i p^i q^{n-i}, & n-i\leqq x<n, \\ =1, & x\geqq n. \end{cases} \qquad \text{(VIII.4,2)}$$

Graficul funcţiei de repartiţie definită de relaţia (VII.4,1) va fi, evident, o funcţie nedescrescătoare cu trepte, în număr egal cu numărul punctelor de discontinuitate.

§ 5. Momente

Calculăm momentele distribuţiei binomiale, folosind teorema sumei şi produsului de variabile independente

$$M(X_1+X_2+\ldots+X_n)=M(X_1)+M(X_2)+\ldots+M(X_n),$$

$$M(X_1\cdot X_2\cdot\ldots\cdot X_n)=M(X_1)\cdot M(X_2)\cdot\ldots\cdot M(X_n).$$

Ştiind că fiecare variabilă aleatorie $X_i\,(i=1, 2, \ldots, n)$ dă naştere distribuţiei

$$X_i\begin{pmatrix}1 & 0\\ p & q\end{pmatrix},$$

putem scrie, ţinînd seama de definiţia momentului (cap. II, § 1)

$$M(X) = M(X_1+X_2+\ldots+X_n) = np,$$

$$M_2(X) = M\left(\sum_i X_i\right)^2 = \sum_i M(X_i^2) + 2\sum_{i\neq j} M(X_i)\,M(X_j) = np +$$

$$+ n(n-1)p^2 = n^2p^2 + npq,$$

$$M_3(X) = M\left(\sum_i X_i\right)^3 = \sum_i M(X_i)^3 + 3\sum_{i\neq j} M(X_i^2)\,M(X_j) +$$

$$+ 6\sum_{i\neq j\neq k} M(X_i)\,M(X_j)\,M(X_k) = np + 3n(n-1)p^2 +$$

$$+ n(n-1)(n-2)p^3 = n^3p^3 + 3n^2p^2q + npq(1-2p),$$

$$M_4(X) = M\left(\sum_i X_i\right)^4 = \sum_i M(X_i^4) + 4\sum_{i\neq j} M(X_i^3)\,M(X_j) +$$

$$+ 6\sum_{i\neq j} M(X_i^2)\,M(X_j^2) + 12\sum_{i\neq j\neq k} M(X_i^2)\,M(X_j)\,M(X_k) +$$

$$+ 24\sum_{i\neq j\neq k\neq l} M(X_i)\,M(X_j)\,M(X_k)\,M(X_l) =$$

$$= np + 4n(n-1)p^2 + 3n(n-1)p^2 + 6n(n-1)(n-2)p^3 +$$

$$+ n(n-1)(n-2)(n-3)p^4 = n^4p^4 + 6n^3p^3q +$$

$$+ n^2p^2(7-18p+11p^2) + np(1-7p+12p^2-6p^3),$$

. . . . . .

$$M_r(X) = n^rp^r + A_1n^{r-1} + A_2n^{r-2} + \ldots + A_{n-1}n,$$

adică, un polinom de gradul r în raport cu n.

Dacă cunoaştem momentele M_r ale distribuţiei binomiale, putem acum să calculăm momentele centrate m_r, momentele factoriale $M_{(r)}$ şi cumulanţii $M_{[r]}$ $(r=1, 2, \ldots, n)$ ale aceleiaşi distribuţii.

Aplicînd relaţia cunoscută

$$m_r = M_r - C_r^1 M_{r-1}\,M + C_r^2 M_{r-2}\,M^2 + \ldots + (-1)^{r-1} C_r^r M^r,$$

obţinem următoarele valori pentru momentele centrate

$$m_1 = 0,$$

$$m_2 = M_2 - M^2 = npq,$$

$$m_3 = M_3 - 3M_1M + 2M^3 = npq\,(q-p),$$

$$m_4 = M_4 - 4M_3M + 6M_2M^2 - 3M^4 = 3n^2p^2q^2 + npq\,(1-6pq),$$

. . . . . .

$$m_{2r} = a_0 n^r + a_1 n^{r-1} + \ldots + a_{r-1} n,$$

$$m_{2r+1} = b_0 n^r + b_1 n^{r-1} + \ldots + b_{r-1} n.$$

Cu alte cuvinte, momentele centrate m_{2r} şi m_{2r+1} ale distribuţiei binomiale sînt polinoame de gradul r în raport cu n.

Momentele centrate se pot obţine mai simplu folosind distribuţia variabilei aleatorii

$$Y = X - np = \sum_{i=1}^{n} (X_i - p), \quad Y_i \begin{pmatrix} q & -p \\ p & q \end{pmatrix}.$$

Atunci, putem scrie

$$m_1 = M(Y) = M\left[\sum_i (X_i - p)\right] = 0,$$

$$m_2 = M_2(Y) = M\left[\sum_i (X_i - p)\right]^2 = \sum_i M(X_i - p)^2 +$$

$$+ 2 \sum_{i \neq j} M(X_i - p)\, M(X_j - p) = npq,$$

$$m_3 = M_3(Y) = M\left[\sum_i (X_i - p)\right]^3 = \sum_i M(X_i - p)^3 +$$

$$+ 3 \sum_{i \neq j} M(X_i - p)^2 M(X_j - p) +$$

$$+ 6 \sum_{(i \neq j \neq k)} M(X_i - p)\, M(X_j - p)\, M(X_k - p) = npq\,(q-p),$$

$$m_4 = M_4(Y) = M\left[\sum_i (X_i - p)\right]^4 = \sum_i M(X_i - p)^4 +$$

$$+ 4 \sum_{i \neq j} M(X_i - p)^3\, M(X_j - p) + 6 \sum_{(i \neq j)} M(X_i - p)^2\, M(X_j - p)^2 +$$

$$+ 12 \sum_{(i \neq j \neq k)} M(X_i - p)^2 M(X_j - p)\, M(X_k - p) +$$

$$+ 24 \sum_{(i \neq j \neq k \neq l)} M(X_i - p)\, M(X_j - p)\, M(X_k - p)\, M(X_l - p) =$$

$$= npq\,(1 - 3pq) + 3n(n-1)\,p^2q^2 = 3n^2p^2q^2 + npq\,(1 - 6pq),$$

. . . . . .

Aplicînd relaţia

$$M_{(r)} = S_r^1 M_1 + S_r^2 M_2 + \ldots + S_r^r M_r ,$$

unde S_r^k $(k=1, 2, \ldots, r)$ sînt numerele lui Stirling de prima speţă, rezultă momentele factoriale

$$M_{(1)} = S_1^1 M_1 = M = np,$$

$$M_{(2)} = S_2^1 M_1 + S_2^2 M_2 = -M + M_2 = n(n-1)p^2,$$

$$M_{(3)} = S_3^1 M_1 + S_3^2 M_2 + S_3^3 M_3 = 2M - 3M_2 + M_3 = n^3p^3 - 3n^2p^3 + 2np^3,$$

$$M_{(4)} = S_4^1 M_1 + S_4^2 M_2 + S_4^3 M_3 + S_4^4 M_4 = \ldots$$

. . . . . .

Pentru calculul cumulanţilor, vom folosi relaţia dintre momente şi cumulanţi (cap. II, § 5) din care deducem

$$M_{[1]} = M_1 = np,$$

$$M_{[2]} = M_2 - M^2 = npq,$$

$$M_{[3]} = M_3 - 3M_2M_1 + 2M^3 = npq(q-p),$$

$$M_{[4]} = M_4 - 4M_3M - 3M_2^2 + 12M_2M^2 - 6M^4 =$$

$$= n^2p^2(7 - 18p + 11p^2) - 4q(2q-p) +$$

$$+ np(1 - 7p + 12p^2 - 6p^3),$$

. . . . . .

§ 6. Funcţia caracteristică

Prin definiţie, *funcţia caracteristică* a distribuţiei binomiale este

$$\varphi(t) = M(e^{\alpha ti}) = \sum_{\alpha=0}^{n} e^{\alpha ti} C_n^\alpha p^\alpha q^{n-\alpha} = \sum C_n^\alpha (pe^{\alpha ti}) q^{n-\alpha} = (pe^{ti} + q)^n .$$

Să considerăm distribuţia variabilei aleatorii x_k, care poate lua valoarea 1 cu probabilitatea p şi valoarea 0 cu probabilitatea $1-p=q$, adică funcţia caracteristică pentru o singură probă

$$X_k \begin{pmatrix} 1 & 0 \\ p & q \end{pmatrix}.$$

Funcţia caracteristică a acestei distribuţii este

$$\varphi_k(t)=pe^{it}+q.$$

Dacă facem n probe independente identice, atunci trebuie să avem în vedere variabila aleatorie $X=\sum_{k=1}^{n} X_k$, X_k fiind variabile independente.

Funcţia caracteristică a variabilei aleatorii X este funcţia caracteristică a distribuţiei binomiale căutate. Ea se deduce aplicînd teorema funcţiei caracteristice sumei de variabile independente.

Obţinem, astfel

$$\varphi(t)=\prod_{k=1}^{n}\varphi_k(t)=(pe^{ti}+q)^n,$$

verificîndu-se astfel valoarea funcţiei caracteristice găsită mai sus. Se ştie (cap. II, § 4) că funcţia caracteristică este o funcţie generatoare de momente

$$\varphi(t)=\sum_{r=0}^{\infty} M_r\frac{(ti)^r}{r!}.$$

Dezvoltînd în serie pe e^{ti}, rezultă următoarea expresie pentru funcţia caracteristică a distribuţiei binomiale

$$\varphi(t)=(pe^{ti}+q)^n=\left[p\left(1+\frac{ti}{1}+\frac{(ti)^2}{2!}+\frac{(ti)^3}{3!}+\ldots+\frac{(ti)^r}{r!}+\ldots\right)+q\right]^n=$$

$$=\left[1+p\frac{ti}{1}+p\frac{(ti)^2}{2!}+p\frac{(ti)^3}{3!}+\ldots+p\frac{(ti)^r}{r!}+\ldots\right]^n=$$

$$=\sum\frac{n!}{a_1!\,a_2!\,a_3!\ldots}(p\,ti)^{a_1}\left(p\frac{(ti)^2}{2!}\right)^{a_2}\left(p\frac{(ti)^3}{3!}\right)^{a_3}\ldots$$

$$(a_1+a_2+a_3+\ldots=n).$$

Efectuînd dezvoltarea sumei, avem

$$\varphi(t)=1+np\frac{ti}{1}+(n^2p^2+npq)\frac{(ti)^2}{2!}+[n^3p^3+3n^2p^2q+$$

$$+npq(1-2p)]\frac{(ti)^3}{3!}+[n^4p^4+6n^3p^3q+n^2p^2(7-18p+11p^2)+$$

$$+np(1-7p+12p^2-6p^3)]\frac{(ti)^4}{4!}+\ldots$$

Regăsim, astfel, aceleaşi valori pentru momentele distribuţiei binomiale, care sînt egale cu coeficienţii lui $\frac{(ti)^r}{r!}$ $(r=1, 2, 3, \ldots)$

$$M_1=np,$$
$$M_2=n^2p^2+npq,$$
$$M_3=n^3p^3+3n^2p^2q+npq\,(1-2p),$$
$$M_4=n^4p^4+6n^3p^3q+n^2p^2\,(7-18p+np^2)+$$
$$+np\,(1-7p+12p^2-6p^3),$$
$$\ldots\ldots\ldots\ldots\ldots$$

Momentele succesive se pot obţine şi pe cale de derivări succesive ale lui $\varphi(t)$.

Folosind procedeul de derivare succesivă în raport cu t a funcţiei caracteristice $\varphi(t)$ şi făcînd apoi $t=0$, se obţine

$$m_1=\frac{1}{i}\frac{d}{dt}[(pe^{qti}+qe^{-pti})^n]_{t=0}=[n\,(pe^{qti}+qe^{-pti})^{n-1}(pqe^{qti}-qpe^{-pti})]=0,$$

$$m_2=\frac{1}{i^2}\frac{d^2}{dt^2}[(pe^{qti}+qe^{-pti})^n]_{t=0}=$$
$$=[n(n-1)\,(pe^{qti}+qe^{-pti})^{n-2}\,(pqe^{qti}-pqe^{-pti})^2+$$
$$+n\,(pe^{qti}+qe^{-pti})^{n-1}\,(pq^2e^{qti}+qp^2e^{-pti})]_{t=0}=npq,$$

$$m_3=\frac{1}{i^3}\cdot\frac{d^3}{dt^3}\,[(pe^{qti}+qe^{-pti})^n]_{t=0}=n\,(n-1)\,(n-2)\,(pe^{qti}+qe^{-pti})^{n-3}\times$$
$$\times(pqe^{qti}-qpe^{-pti})^3+2n\,(n-1)\,(pe^{qti}+qe^{-pti})^{n-2}(pqe^{qti}-qpe^{-pti})\times$$
$$\times(pq^2e^{qti}+qp^2e^{-pti})+n\,(n-1)\,(pe^{qti}+qe^{-pti})^{n-2}\,(pq^2e^{qti}+qp^2e^{-pti})\times$$
$$\times(pqe^{qti}-qpe^{-pti})+n\,(pe^{qti}+qe^{-pti})^{n-1}\,(pq^3e^{qti}-qp^3e^{-pti})]_{t=0}=$$
$$=npq\,(q-p).$$

Momentele factoriale ale distribuţiei binomiale se pot calcula imediat, aplicînd relaţia

$$M_{(r)}=\frac{d^r}{dt^r}\,[\Phi(t)]_{t=1},$$

unde $\Phi(t)=\int_{-\infty}^{\infty} t^x\,dF(x)$ este *funcţia generatoare a momentelor factoriale.*

In cazul distribuţiei binomiale, avem

$$\Phi(t)=(pt+q)^n. \qquad \text{(VII.6,1)}$$

Această soluţie se obţine observînd că funcţia $\Phi(t)$ este $M(t^x)$, momentul lui t^x.

Prin urmare, vom putea scrie

$$\Phi(t)=M(t^x)=\sum_{\alpha=0}^{n} C_n^\alpha p^\alpha q^{n-\alpha} t^\alpha=\sum_{\alpha=0}^{n} C_n^\alpha (pt)^\alpha q^{n-\alpha}=(pt+q)^n.$$

Relaţia (VII.6,1) se mai poate stabili considerînd $M\left(t^{\sum_{i=1}^{n} x_i}\right)$, unde X_i sînt variabile independente şi pot lua fiecare valoarea 1 cu probabilitatea p şi valoarea 0 cu probabilitatea q.

Atunci, vom scrie

$$\Phi(t)=M(t^{x_1+x_2+\cdots+x_n})=M(t^{x_1})\cdot M(t^{x_2})\ \ldots\ M(t^{x_n}). \qquad \text{(VII.6,2)}$$

Deoarece

$$M(t^{x_i})=pt+qt^0=pt+q \qquad (i=1,\ 2,\ \ldots,\ n)$$

rezultă, în baza relaţiei (VII.6,2),

$$\Phi(t)=(pt+q)^n.$$

Luînd derivatele în raport cu t şi făcînd pe $t=1$, vom deduce momentele factoriale ale distribuţiei binomiale

$$M_{(1)}=\frac{d}{dt}\left[(pt+q)^n\right]_{t=1}=\left[np\,(pt+q)^{n-1}\right]_{t=1}=np,$$

$$M_{(2)}=\frac{d^2}{dt^2}\left[(pt+q)^n\right]_{t=1}=\left[n(n-1)\,p^2\,(pt+q)^{n-2}\right]_{t=1}=n(n-1)\,p^2,$$

$$M_{(3)} = \frac{d^3}{dt^3}[(pt+q)^n]_{t=1} = [n(n-1)(n-2)p^3(pt+q)^{n-3}]_{t=1} =$$

$$= n(n-1)(n-2)p^3,$$

. . . . . .

$$M_{(r)} = \frac{d^r}{dt^r}[(pt+q)^n]_{t=1} = [n(n-1)\ldots(n-r+1)p^r(pt+q)^{n-r}]_{t=1} =$$

$$= n(n-1)\ldots(n-r+1)p^r.$$

§ 7. Expresia asimptotică a probabilității $P_{n,\alpha}$

Ne propunem să calculăm valoarea probabilității

$$P_{n,\alpha} = C_n^\alpha p^\alpha q^{n-\alpha} = \frac{n!}{\alpha!(n-\alpha)!} p^\alpha q^{n-\alpha} \qquad \text{(VII.7,1)}$$

pentru valorile mari ale lui n.

Pentru valorile mici ale lui n, folosirea în aplicații al acestei probabilități este, în general, lipsită de interes. Aplicațiile ei cele mai importante se întîlnesc pentru valorile mari ale lui n, adică în cazul colectivelor cu volum mare.

Insă, pentru valorile mari date lui n, un calcul direct al lui $P_{n,\alpha}$ este practic imposibil, din cauza factorialelor care necesită pentru determinarea lor numerică un număr mare de operații.

Această greutate se înlătură folosind formula asimptotică a lui Moivre-Stirling

$$n! = n^n e^{-n} \sqrt{2\pi n}\,(1+O(n)),$$

unde $O(n)$ tinde către zero, cînd n crește necontenit.

Calculele expuse în cap. II, § 7 arată că $O(n)$ este mai mic în valoare absolută decît $\frac{1}{10n}$. Pentru aplicațiile noastre numerice, influența lui $O(n)$ este neglijabilă și de aceea vom folosi formula lui Moivre-Stirling fără restul $O(n)$. Să introducem abaterea frecvenței α de la valoarea ei medie np, adică

$$l = \alpha - np.$$

Inlocuind pe α cu $l+np$ și notînd pe $P_{n,\alpha}$ cu $P_{n,l}$, expresia (VII.7,1) devine

$$P_{n,l} = \frac{n!}{(l+np)!\,(nq-l)!} p^{l+np} \cdot q^{nq-l}. \qquad \text{(VII.7,2)}$$

Folosind acum formula lui Moivre-Stirling şi efectuînd simplificările necesare, ajungem la relaţia

$$P_{n,\,l}=\frac{1}{\sqrt{2\pi npq}\left(1+\frac{l}{np}\right)^{l+np+\frac{1}{2}}\cdot\left(1-\frac{l}{nq}\right)^{nq-l+\frac{1}{2}}}.$$

Pentru prescurtare, să punem

$$N=\left(1+\frac{l}{np}\right)^{np+l+\frac{1}{2}}\cdot\left(1-\frac{l}{nq}\right)^{nq-l+\frac{1}{2}}.$$

Aplicînd logaritmii

$$\ln N=\left(np+l+\frac{1}{2}\right)\ln\left(1+\frac{l}{np}\right)+\left(nq-l+\frac{1}{2}\right)\ln\left(1-\frac{l}{nq}\right). \qquad \text{(VII.7,3)}$$

Evaluăm asimptotic pe $P_{n,l}$, pentru valorile lui $|l|<A\sqrt{n}$, unde A este un număr pozitiv dat.
In acest caz, pentru valorile mari ale lui n avem

$$\frac{|l|}{np}<1, \quad \frac{|l|}{nq}<1.$$

Putem, deci, face dezvoltările în serie

$$\ln\left(1+\frac{l}{np}\right)=\frac{l}{np}-\frac{l^2}{2n^2p^2}+\frac{l^3}{3n^3p^3}-\cdots,$$

$$\ln\left(1-\frac{l}{nq}\right)=-\frac{l}{nq}-\frac{l^2}{2n^2q^2}-\frac{l^3}{3n^3q^3}-\cdots \qquad \text{(VII.7,4)}$$

Inlocuind în (VII.7,3), expresiile lui $\ln\left(1+\frac{l}{np}\right)$ şi $\ln\left(1-\frac{l}{nq}\right)$ date de dezvoltările în serie (VII.7,4) şi ţinînd seama că $|l|<A\sqrt{n}$, obţinem următoarea valoare asimptotică

$$\ln N=\frac{l^2}{2npq}+O\left(\frac{1}{\sqrt{n}}\right).$$

Scanned with OKEN Scanner

Atunci, pentru valorile lui n suficient de mari, vom putea scrie

$$N \simeq e^{\frac{l^2}{2npq}}$$

şi, prin urmare, expresia (VII.7,2) a probabilităţii $P_{n,l}$ devine

$$P_{n,l} = \frac{1}{\sqrt{2\pi npq}} e^{-\frac{l^2}{2npq}}. \qquad \text{(VII.7,5)}$$

Legea de probabilitate (VII.7,5) este o *lege normală de probabilitate.*

Observînd că $\sqrt{npq}$ este abaterea medie pătratică σ, expresia (VII.7,5) se mai poate scrie

$$P_{n,l} = \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-\frac{l^2}{2\sigma^2}}. \qquad \text{(VII.7,6)}$$

Acest rezultat dovedeşte utilitatea stabilirii formulei asimptotice a probabilităţii $P_{n,l}$ cu ajutorul formulei lui Moivre-Stirling, deoarece prin acest procedeu se ajunge nu numai la posibilitatea evaluării acestei probabilităţi cu mai puţine calcule numerice, dar şi la o formulă care prezintă importanţă teoretică.

Examinînd expresia asimptotică (VII.7,6) a probabilităţii $P_{n,l}$, s-ar putea trage concluzia că este inacceptabilă.

In adevăr, în baza acestei formule, rezultă

$$P_{n,l} = P_{n,-l},$$

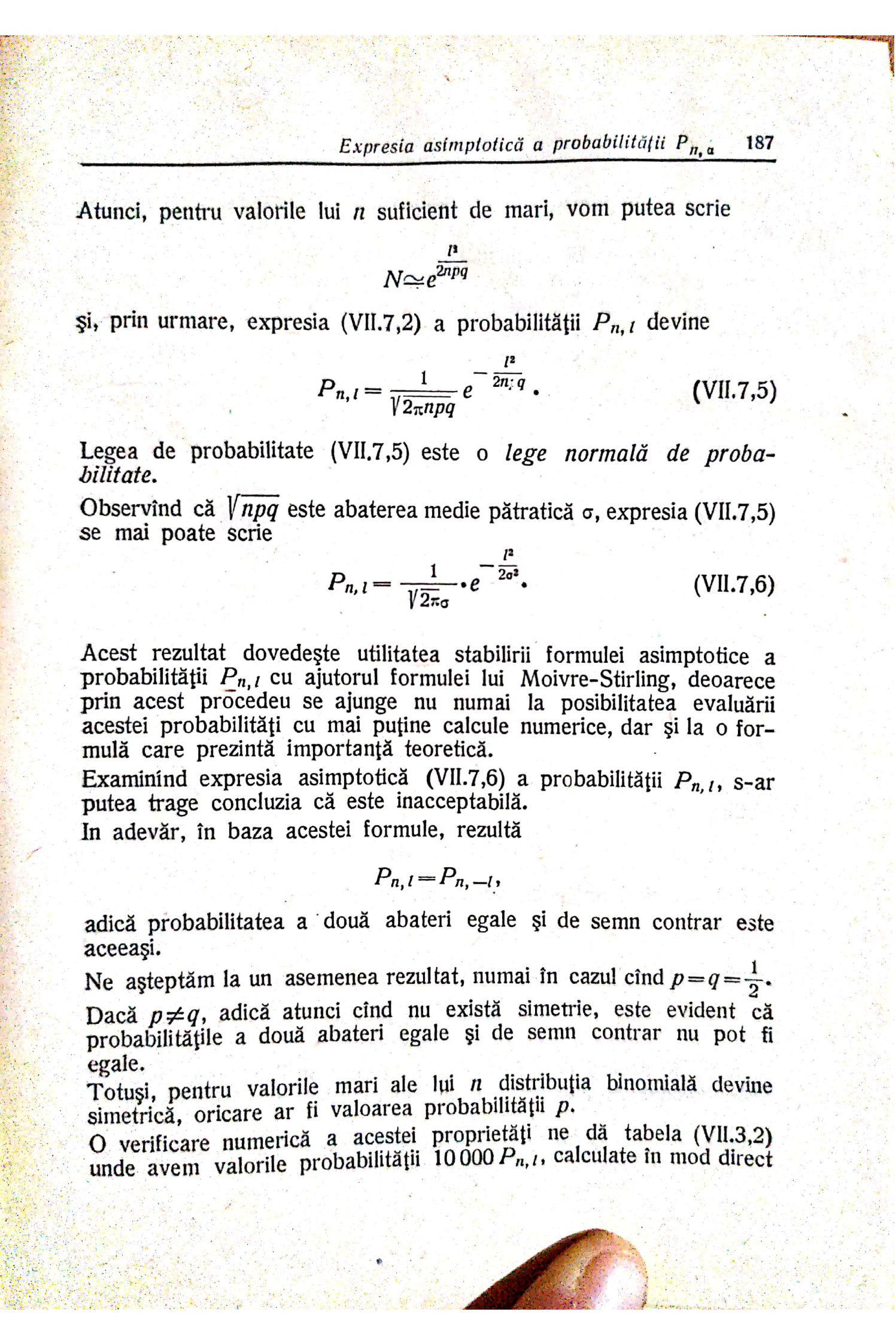
adică probabilitatea a două abateri egale şi de semn contrar este aceeaşi.

Ne aşteptăm la un asemenea rezultat, numai în cazul cînd $p = q = \frac{1}{2}$.

Dacă $p \neq q$, adică atunci cînd nu există simetrie, este evident că probabilităţile a două abateri egale şi de semn contrar nu pot fi egale.

Totuşi, pentru valorile mari ale lui n distribuţia binomială devine simetrică, oricare ar fi valoarea probabilităţii p.

O verificare numerică a acestei proprietăţi ne dă tabela (VII.3,2) unde avem valorile probabilităţii $10\,000\,P_{n,l}$, calculate în mod direct



pentru $p=0,1$ şi $q=0,9$ deci într-un caz în care, în mod sensibil, avem $p \neq q$.
După cum observăm, pentru $n=5$, distribuţia este asimetrică. Din cauză că p este foarte mic faţă de q, probabilităţile ca evenimentul să nu se verifice sînt mai mari decît probabilitatea ca el să se verifice, adică, probabilitatea de a obţine de cinci ori la rînd valoarea zero este 0,590 5, care este cea mai mare faţă de celelalte probabilităţii.
Pe măsură ce n creşte tabela probabilităţilor începe să devină simetrică. Aşa, de exemplu, pentru $n=100$, constatăm o simetrie destul de pronunţată faţă de valoarea medie $np=10$, căreia îi corespunde probabilitatea maximă 0,131 9.
In concluzie, calculul direct confirmă expresia (VII.7,6) a probabilităţii $P_{n,l}$, care nu se aplică decît pentru valorile mari ale lui n.

§ 8. Distribuţia limită

Formula (VII.7,5) nu rezolvă problema distribuţiei limită a schemei binomiale. Trebuie să se demonstreze că, într-adevăr, pentru $n \to \infty$, legea de repartiţie a distribuţiei binomiale tinde către legea de repartiţie a legii normale. Cu alte cuvinte, trebuie să dovedim că formula (VII.7,5), stabilită pentru o anumită valoare a abaterii l, poate fi folosită prin însumare şi într-un interval de variaţie a lui l.
Pornim din nou de la probabilitatea ca din n extracţii succesive să obţinem α bile albe

$$P_{n,\alpha} = \frac{n!}{\alpha!\,(n-\alpha)!}\, p^{\alpha} \cdot q^{n-\alpha}.$$

Folosim acelaşi procedeu de calcul, notăm însă

$$\alpha = np + x\sqrt{2npq} \qquad \text{(VII.8,1)}$$

pentru a pune în evidenţă abaterea relativă

$$x = \frac{\alpha - np}{\sqrt{2npq}}.$$

Aplicînd în acelaşi mod formula lui Moivre-Stirling, vom arăta că probabilitatea $P_{n,\alpha}$ se poate scrie sub forma

$$P_{n,\alpha} = \frac{e^{-x^2}}{\sqrt{2npq}}\,[1 + h(n,\alpha)],$$

valabilă în ipoteza că x are o valoare cuprinsă într-un interval finit, independent de n.
Formula (VII.8,1) diferă de formula (VII.7,6) prin faptul că se introduce termenul $h(n, \alpha)$ care tinde uniform către zero, cînd n şi α tind către infinit, fiind în intervalul finit sus-menţionat.
Să notăm cu

$$P_n(-\lambda, \lambda)$$

probabilitatea inegalităţilor

$$-\lambda\sqrt{2npq} < \alpha - np < \lambda\sqrt{2npq}, \qquad \text{(VII.8,2)}$$

unde α ia valori întregi, de la 0 la n, verificînd inegalităţile (VII.8,2).
Fie ξ şi η respectiv cea mai mică şi cea mai mare valoare a lui α care verifică aceste inegalităţi.
In baza proprietăţii de adunare a probabilităţilor evenimentelor incompatibile obţinem

$$P_n(-\lambda, \lambda) = \sum_{\alpha=\xi}^{\eta} P_{n,\alpha}.$$

Să mai observăm că din egalitatea (VII.8.1) reiese că, pentru fiecare valoare a lui α, rezultă o valoare bine determinată a lui x, pe care să o notăm cu $x(\alpha)$. Ţinînd seama şi de (VII.8,2), obţinem

$$P_n(-\lambda, \lambda) = \sum_{\xi}^{\eta} \frac{e^{-x^2(\alpha)}}{\sqrt{2npq}} [1 + h(n, \alpha)].$$

Fie $M(n)$ cea mai mare dintre valorile absolute ale valorilor $h(n, \alpha)$, pentru α cuprins între ξ şi η inclusiv.
Formula anterioară se poate scrie

$$P_n(-\lambda, \lambda) = S_n[1 + h(n)],$$

unde

$$h(n) < M(n)$$

iar

$$S_n = \sum_{\alpha=\xi}^{\eta} \frac{e^{-x^2(\alpha)}}{\sqrt{2\pi npq}}.$$

Ţinînd seama de valorile lui $x(\alpha)$, S_n se mai poate scrie

$$S_n = \frac{1}{\sqrt{\pi}} \sum_{\alpha=\xi}^{\eta} e^{-x^2(\alpha)} [x(\alpha+1) - x(\alpha)].$$

Scanned with OKEN Scanner

Prin urmare, trecînd la limită

$$\lim_{n\to\infty} S_n = \frac{1}{\sqrt{\pi}} \int_{-\lambda}^{\lambda} e^{-x^2} dx,$$

deoarece $x(\alpha)$, pentru $\alpha=\xi$ şi $\alpha=\eta$, ia respectiv valorile $-\lambda, \lambda$.
Pe de altă parte, am arătat, că dacă λ ia o valoare finită, $M(n)$ converge uniform către zero.
Proprietatea de limită este deci dovedită.
Formulele stabilite mai sus servesc la numeroase aplicaţii.
Am arătat astfel că

$$\lim_{n\to\infty} P_n\left(-\xi\sqrt{2npq} < \alpha - np < \xi\sqrt{2npq}\right) = \frac{2}{\sqrt{\pi}} \int_0^{\xi} e^{-x^2} dx.$$

Deci, pentru a calcula probabilitatea inegalităţii

$$-\xi\sqrt{2npq} < \alpha - np < \xi\sqrt{2npq}$$

pentru valorile mari ale lui n, putem să folosim funcţia $H(\xi)$ determinată prin egalitatea

$$H(\xi) = \frac{2}{\sqrt{\pi}} \int_0^{\xi} e^{-x^2} dx.$$

Valoarea acestei probabilităţi este egală cu aria mărginită de axa Ox, de dreptele $x=-\xi$, $x=\xi$ şi de curba

$$y = e^{-x^2}.$$

In cazul abaterilor absolute, se ajunge la formule de tipul

$$P_n(-\lambda, \lambda) = \frac{h}{\sqrt{\pi}} \int_0^{\lambda} e^{-h^2x^2} dx,$$

unde

$$h = \frac{1}{\sqrt{2npq}}.$$

De data aceasta, valoarea probabilităţii $P_n(-\lambda, \lambda)$ este dată de aria delimitată de axa Ox, de dreptele $x=-\lambda$, $x=\lambda$ şi de curba

$$y = \frac{1}{\sqrt{\pi}} e^{-h^2x^2}. \qquad \text{(VII.8,3)}$$

Curba de probabilitate (VII.8,3) depinde de un singur parametru h. Ea este simetrică față de axa Oy, are un singur maxim pe această axă în punctul $\left(0, \frac{h}{\sqrt{\pi}}\right)$ și se apropie asimptotic de axa Ox.

Pentru a arăta cum se aplică formulele de mai sus, dăm exemplele următoare.

Exemplul 1. La extragerea repetată a 10 000 bile dintr-o urnă cu același număr de bile albe și negre, s-au obținut 5 050 bile albe și 4 950 bile negre. Să se determine probabilitatea ca la repetarea experienței să nu obținem o abatere mai mare.

Probabilitatea căutată este dată de formula

$$P[-\xi\sqrt{2npq} < \alpha - np < \xi\sqrt{2npq}] \approx H(\xi).$$

Din datele problemei rezultă

$$\alpha = 5\,050; \quad n = 10\,000; \quad p = \frac{1}{2}, \quad \text{deci} \quad \alpha - np = 50,$$

$$\xi = \frac{\alpha - np}{\sqrt{2npq}} = \frac{50}{\sqrt{2 \cdot 10\,000 \cdot \frac{1}{4}}} = \frac{50}{10\sqrt{50}} = 0{,}70.$$

Din tabela care dă valorile funcției $H(\xi)$ aflăm valoarea

$$H(0{,}70) = 0{,}67.$$

Prin urmare probabilitatea căutată este 0,67.

Exemplul 2. Știind că din 1 000 piese 30 sînt date la rebut, să se determine probabilitatea ca, din 6 000 piese fabricate, numărul celor bune să fie cuprins între 5 780 și 5 860.

$$n = 6\,000; \quad p = \frac{97}{100}; \quad q = \frac{3}{100}; \quad \alpha - np = 40;$$

$$\xi = \frac{\alpha - np}{\sqrt{2npq}} = \frac{40}{\sqrt{2 \times 6\,000 \times \frac{97}{100} \times \frac{3}{100}}} \approx 2{,}15.$$

Probabilitatea căutată este deci

$$P = H(\xi) = H(2{,}15) = 0{,}997\,64.$$

§ 9. O altă demonstrație a teoremei de limită

Legea limită a distribuției binomiale se poate stabili în mod simplu, folosind funcția caracteristică a variabilei reduse

$$\frac{\alpha - np}{\sqrt{npq}}.$$

Potrivit definiției funcției caracteristice, avem în cazul variabilei reduse

$$\varphi_n(t) = M\left[e^{\frac{\alpha-np}{\sqrt{npq}}ti}\right] = e^{-np\frac{t}{\sqrt{npq}}} M\left(e^{\alpha\frac{t}{\sqrt{npq}}i}\right) =$$

$$= \left(pe^{q\frac{t}{\sqrt{npq}}i} + qe^{-p\frac{t}{\sqrt{npq}}i}\right)^n. \qquad \text{(VII. 9,1)}$$

Folosind formula de dezvoltare

$$e^{xi} = 1 + \frac{xi}{1!} + \frac{(xi)^2}{2!} + \theta\frac{x^3}{3}, \quad |\theta| < 1$$

obținem pentru $\varphi_n(t)$, din expresia (VII. 9,1)

$$\varphi_n(t) = \left[p\left(1 + \frac{qti}{\sqrt{npq}} + \frac{(qti)^2}{2!\,npq} + \theta_1\frac{(qt)^3}{3!(npq)^{\frac{3}{2}}}\right) + q\left(1 - \frac{pti}{\sqrt{npq}} + \right.\right.$$

$$\left.\left. + \frac{(pti)^2}{2!\,npq} + \theta_2\frac{(pt)^3}{3!(npq)^{\frac{3}{2}}}\right)\right]^n$$

sau, efectuînd calculele,

$$\varphi_n(t) = \left[1 - \frac{t^2}{2n} + O\left(n^{-\frac{3}{2}}\right)\right]^n,$$

unde $O\left(n^{-\frac{3}{2}}\right)$ este o funcție de n, asimptotică de ordinul $n^{-\frac{3}{2}}$. Pentru determinarea limitei lui $\varphi_n(t)$ pentru $n \to \infty$, vom aplica logaritmii în relația de mai sus

$$\ln \varphi_n(t) = n \ln\left[1 - \frac{t^2}{2n} + O\left(n^{-\frac{3}{2}}\right)\right] = \frac{\ln\left[1 - \frac{t^2}{2n} + O\left(n^{-\frac{3}{2}}\right)\right]}{\frac{1}{n}}.$$

Notînd $\left(-\frac{t^2}{2}+nO\left(n^{-\frac{3}{2}}\right)\right)$ cu u, logaritmul de mai sus devine

$$\ln \varphi_n(t)=\frac{\ln\left(1+\frac{u}{n}\right)}{\frac{1}{n}}=\frac{\ln(1+\alpha u)}{\alpha},$$

dacă punem $\frac{1}{n}=\alpha$.

Rezultă

$$\lim_{n\to\infty}\ln\varphi_n(t)=\lim_{\alpha\to 0}\frac{u}{\frac{1+\alpha u}{1}}=-\frac{t^2}{2}.$$

Prin urmare,

$$\lim_{n\to\infty}\varphi_n(t)=e^{-\frac{t^2}{2}},$$

adică, funcția caracteristică $\varphi_n(t)$ a distribuției binomiale tinde, cu $n\to\infty$, către funcția caracteristică cunoscută $e^{-\frac{t^2}{2}}$ a legii normale. După cum știm, dacă un șir de funcții caracteristice tinde către funcția caracteristică a legii normale, legile de repartiție respective tind și ele către legea de repartiție a legii normale.

Teorema este deci demonstrată.

Mai notăm că, în cazul distribuției binomiale coeficienții de asimetrie și exces sînt

$$\gamma_1=\frac{q-p}{\sqrt{npq}},\quad \gamma_2=\frac{1-6pq}{npq}.$$

Să observăm că pentru $n\to\infty$, coeficienții γ_1 și γ_2 tind fiecare către zero, ceea ce este un indiciu că, pentru n crescînd, distribuția binomială se apropie de simetrie și normalitate.

Capitolul VIII

Distribuţia multinomială

§ 1. Consideraţii generale

In capitolul precedent am studiat distribuţia binomială, care dă naştere la o succesiune de variabile aleatorii independente

$$X_1,\ X_2,\ldots,\ X_s,\ldots,$$

unde fiecare dintre variabilele X_i $(i=1,\ 2,\ldots)$ poate lua numai valorile 1 sau 0, cu probabilităţile respective p şi q $(p+q=1)$.
Generalizarea acestei distribuţii se impune în mod natural, considerînd mai multe stări în loc de două.
Distribuţia multinomială constituie un exemplu simplu de vector aleator.
Să admitem că valorile $a_1,\ a_2,\ldots,\ a_m$ reprezintă coordonatele unui vector A într-un spaţiu liniar L_m, în care luăm ca bază vectorii independenţi

$$e_1 = (1,\ 0,\ 0,\ldots,0,\ 0),$$
$$e_2 = (0,\ 1,\ 0,\ldots,0,\ 0),$$
$$\cdot\quad\cdot\quad\cdot\quad\cdot\quad\cdot\quad\cdot\quad\cdot$$
$$e_m = (0,\ 0,\ 0,\ldots,0,\ 1).$$

Să considerăm variabilele vectoriale aleatorii independente X_k $(k=1,\ 2,\ldots,\ n)$ care pot lua fiecare vectorii $e_1,\ e_2,\ldots,\ e_m$ cu probabilităţile respective $p_1,\ p_2,\ldots,\ p_m$ $\left(p_i \geqq 0,\ \sum_{i=1}^{m} p_i = 1\right)$.
Distribuţia sumei $X_1 + X_2 +\ldots+ X_n$ poartă numele de *distribuţie multinomială*.
Egalitatea

$$X_1 + X_2 +\ldots+ X_n = A$$

înseamnă că în cele n experienţe independente, realizate de variabilele $X_1,\ X_2,\ldots,\ X_n$ vectorul e_1 a fost luat de a_1 ori, vectorul e_2 de a_2 etc.
In aplicaţiile statisticii matematice se întîlnesc multe distribuţii de acest gen.

Scanned with OKEN Scanner

Exemplul 1. Să aruncăm un zar de zece ori. Fiecare aruncare cu zarul este o probă independentă, care constă în apariția unui număr de puncte egal cu 1, 2, 3, 4, 5 sau 6.

Sistemului celor zece probe corespunde suma variabilelor aleatorii independente X_k, fiecare din aceste variabile putînd lua valorile 1, 2, ..., 6 cu probabilitățile egale fiecare cu o șesime.

Avem deci, distribuția

$$X_k \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \end{pmatrix} \qquad (k=1, 2, \ldots, 10).$$

Fiecare aruncare cu zarul mai poate fi privită ca o probă care constă în apariția uneia dintre fețe și, evident, în neapariția celorlalte cinci fețe ale zarului.

In acest caz, vom avea de considerat o succesiune de variabile vectoriale aleatorii independente X_k, fiecare dintre ele putînd fi unul dintre vectorii

$$\xi_1 (1, 0, 0, \ldots, 0), \quad \xi_2 (0, 1, 0, \ldots, 0), \ldots, \quad \xi_6 (0, 0, 0, \ldots, 1),$$

cu probabilitățile respective

$$p_1 = \frac{1}{6}, \quad p_2 = \frac{1}{6}, \ldots, \quad p_6 = \frac{1}{6}.$$

Avem deci distribuțiile individuale pentru fiecare probă

$$X_k \begin{pmatrix} \xi_1 (1, 0, \ldots, 0), & \xi_2 (0, 1, \ldots, 0), \ldots, & \xi_6 (0, \ldots, 0, 1) \\ \frac{1}{6}, & \frac{1}{6}, & \frac{1}{6} \end{pmatrix} (k=1, 2, \ldots, 10).$$

Exemplul 2. Am arătat că mortalitatea unui grup omogen de N persoane de aceeași vîrstă este un fenomen statistic care se desfășoară după schema distribuției binomiale.

In cazul în care intervine și factorul invalidității la un grup de N persoane active de aceeași vîrstă, avînd toate, pe cît este posibil, aceleași condiții de viață, ne vom găsi în fața unei distribuții multinomiale cu trei stări.

Intr-adevăr din cele N persoane active de aceeași vîrstă, pot să fie, după trecerea unui an, A persoane în stare de activitate, B persoane în stare de invaliditate și C persoane decedate cu probabilitățile respective P, Q și R.

In modul acesta, avem variabila aleatorie X, care dă naștere distribuției

$$X \begin{pmatrix} A, & B, & C \\ P, & Q, & R \end{pmatrix} \qquad (P+Q+R=1).$$

O singură persoană, din grupul de N persoane considerate, poate să se găsească după trecerea unui an într-una din cele trei stări și să nu se găsească în celelalte două stări menționate.

Atunci, vom avea șirul de variabile aleatorii independente

$$X_k\begin{pmatrix}\xi_1(1,0,0), & \xi_2(0,1,0), & \xi_3(0,0,1)\\ p, & q, & r\end{pmatrix} \qquad \begin{matrix}(k=1,2,\ldots,N),\\ (p+q+r=1),\end{matrix}$$

unde p, q, r sînt probabilități cunoscute.

Prin însumarea variabilelor X_k $(k=1, 2, \ldots, N)$, ajungem la variabila X și la determinarea probabilităților P, Q, R în funcție de p, q, r și N.

Din considerațiile făcute pînă aici, rezultă ușor o schemă generală care să servească drept model pentru fenomene statistice, similare cu cele din exemplele arătate.

Se ne imaginăm o urnă care conține n_1 bile de o culoare, n_2 bile de altă culoare etc. Din această urnă facem o succesiune de m extracții, punînd după fiecare extracție, bila extrasă înapoi în urnă.

Aceste m extracții succesive dau m probe independente, deci un șir de variabile aleatorii independente, unde fiecare dintre aceste variabile poate lua valorile vectoriale

$$\xi_1(1, 0, \ldots, 0), \quad \xi_2(0, 1, 0, \ldots, 0), \ldots, \quad \xi_m(0, 0, \ldots, 0, 1)$$

cu probabilitățile

$$p_1=\frac{n_1}{\sum_{i=1}^{m} n_i}, \quad p_2=\frac{n_2}{\sum_{i=1}^{m} n_i}, \ldots, \quad p_m=\frac{n_m}{\sum_{i=1}^{m} n_i}.$$

Avem astfel *schema lui J. Bernoulli cu mai multe stări.*

§ 2. Determinarea probabilității ca în n probe succesive să obținem vectorii e_i respectiv de α_i ori

Teoremă. *Fie succesiunea de variabile aleatorii independente*

$$X_1, X_2, \ldots, X_s, \ldots,$$

unde fiecare dintre variabilele aleatorii X_k poate lua vectorii

$$e_1, e_2, \ldots, e_m$$

cu probabilitățile

$$p_1, p_2, \ldots, p_m \qquad \left(\sum_{k=1}^{m} p_k=1\right).$$

Atunci, probabilitatea ca să obținem

$$e_1 \text{ de } \alpha_1 \text{ ori}, \quad e_2 \text{ de } \alpha_2 \text{ ori}, \ldots, \quad e_m \text{ de } \alpha_m \text{ ori}$$

este

$$P_n(\alpha_1, \alpha_2, \ldots, \alpha_m) = \frac{n!}{\alpha_1!\,\alpha_2!\ldots\alpha_m!}\, p_1^{\alpha_1} \cdot p_2^{\alpha_2} \ldots p_m^{\alpha_m} \quad \left(\sum_{i=1}^{m} \alpha_i = n\right).$$

Vom stabili formula pentru $m=3$.

In acest caz, fiecare variabilă X_k poate lua numai valorile

$$e_1,\ e_2,\ e_3$$

cu probabilităţile

$$p_1,\ p_2,\ p_3 \qquad (p_1+p_2+p_3=1).$$

Ne propunem, deci, să determinăm probabilitatea $P_n(\alpha_1, \alpha_2, \alpha_3)$ ca să obţinem vectorul e_1 de α_1 ori, vectorul e_2 de α_2 ori şi vectorul e_3 de α_3 ori $(\alpha_1+\alpha_2+\alpha_3=n)$.

Să presupunem că obţinem aceste valori în ordinea următoare

$$\underbrace{e_1,\ e_1,\ldots,\ e_1}_{\text{de } \alpha_1 \text{ ori}},\quad \underbrace{e_2,\ e_2,\ldots,\ e_2}_{\text{de } \alpha_2 \text{ ori}},\quad \underbrace{e_3,\ e_3,\ldots,\ e_3}_{\text{de } \alpha_3 \text{ ori}}.$$

Atunci, probabilitatea de a avea această succesiune, în baza teoremei probabilităţilor compuse, va fi

$$p_1^{\alpha_1} \cdot p_2^{\alpha_2} \cdot p_3^{\alpha_3}.$$

Deoarece, succesiunea valorilor e_1 de α_1 ori, e_2 de α_2 ori şi e_3 de α_3 ori poate fi oricum, vor rezulta, evident, atîtea produse $p_1^{\alpha_1} \cdot p_2^{\alpha_2} \cdot p_3^{\alpha_3}$ cîte combinări posibile putem face din cele $\alpha_1+\alpha_2+\alpha_3=n$ elemente, dintre care α_1 identice cu e_1, α_2 identice cu e_2 şi α_3 identice cu e_3. Numărul acestor combinări este egal cu

$$C_n^{\alpha_1} \cdot C_{n-\alpha_1}^{\alpha_2} \cdot C_{n-\alpha_1-\alpha_2}^{\alpha_3} = \frac{n!}{\alpha_1!\,(n-\alpha_1)!} \cdot \frac{(n-\alpha_1)!}{\alpha_2!\,(n-\alpha_1-\alpha_2)!} = \frac{n!}{\alpha_1!\,\alpha_2!\,\alpha_3!}, \tag{VIII.2,1}$$

întrucît $n-\alpha_1-\alpha_2=\alpha_3$ şi deci $C_{n-\alpha_1-\alpha_2}^{\alpha_3}=1$.

La relaţia (VIII.2,1) ajungem considerînd numărul combinărilor de n elemente luate cîte α_1, care asociate cu numărul combinărilor de $n-\alpha_1$ elemente rămase, luate cîte α_2, fac $C_n^{\alpha_1} \cdot C_{n-\alpha_1}^{\alpha_2}$. La acest număr trebuie acum să asociem numărul de combinări de $n-(\alpha_1+\alpha_2)$ elemente rămase, luate cîte α_3 şi obţinem astfel $C_n^{\alpha_1} \cdot C_{n-\alpha_1}^{\alpha_2} \cdot C_{n-\alpha_1-\alpha_2}^{\alpha_3}$.

Scanned with OKEN Scanner

Probabilitatea căutată este deci

$$P_n(\alpha_1,\ \alpha_2,\ \alpha_3)=\frac{n!}{\alpha_1!\,\alpha_2!\,\alpha_3!}p_1^{\alpha_1}\cdot p_2^{\alpha_2}\cdot p_3^{\alpha_3}. \qquad \text{(VIII.2,2)}$$

In acelaşi mod se stabileşte formula pentru $m>3$.

Avem

$$\sum P_n(\alpha_1,\ \alpha_2,\ \dots,\ \alpha_m)=1,$$

în suma de mai sus α_i luînd toate valorile posibile, astfel ca $\alpha_i \geqq 0,\ \sum_{i=1}^{m}\alpha_i=n.$

Aceasta este evident, deoarece

$$(p_1+p_2+\dots+p_m)^n=\sum_{\alpha_i}\frac{n!}{\alpha_1!\,\alpha_2!\,\dots\,\alpha_m!}p_1^{\alpha_1}\cdot p_2^{\alpha_2}\cdot\dots\cdot p_m^{\alpha_m},$$

suma fiind extinsă la toţi indicii $\alpha_i \geqq 0,\ \sum_{i=1}^{m}\alpha_i=n,$

iar

$$p_1+p_2+\dots+p_m=1.$$

§ 3. Valoarea asimptotică a probabilităţii $P_n(\alpha_1, \alpha_2, .., \alpha_m)$

Considerînd variabilele aleatorii independente

$$X_k\begin{pmatrix}e_1, & e_2, \dots, e_m\\ p_1, & p_2, \dots, p_m\end{pmatrix} \qquad \begin{matrix}(k=1,\ 2, \dots, n),\\ \left(\sum_{i=1}^{m}p_i=1\right),\end{matrix}$$

am arătat că probabilitatea de a obţine

e_1 de α_1 ori, e_2 de α_2 ori, ..., e_m de α_m ori

este

$$P_n(\alpha_1,\ \alpha_2, \dots,\ \alpha_m)=\frac{n!}{\alpha_1!\,\alpha_2!\,\dots\,\alpha_m!}p_1^{\alpha_1}\cdot p_2^{\alpha_2}\dots p_m^{\alpha_m}.$$

Să introducem variabila redusă

$$\xi_i=\frac{\alpha_i-np_i}{\sqrt{np_iq_i}} \qquad \begin{matrix}(i=1,\ 2, \dots, m),\\ (q_i=1-p_i).\end{matrix}$$

Teoremă. *Oricare ar fi* $a<\xi_i<b$, *unde* a *și* b *sînt două constante, există convergența*

$$n^{\frac{m-1}{2}} P_n(\alpha_1, \alpha_2, \dots, \alpha_m) : \frac{e^{-\frac{1}{2}\sum_{i=1}^{m} q_i \xi_i^2}}{(2\pi)^{\frac{m-1}{2}} \sqrt{p_1 \cdot p_2 \dots p_m}} \to 1, \qquad (n \to \infty).$$

Pentru demonstrarea teoremei, folosim formula lui Stirling

$$n! = \sqrt{2\pi n}\, e^{-n} e^{\theta_n}, \qquad |\theta_n| \leq \frac{1}{12n}.$$

Putem deci scrie

$$P_n(\alpha_1, \alpha_2, \dots, \alpha_m) = \frac{n!}{\alpha_1!\, \alpha_2! \dots \alpha_m!} p_1^{\alpha_1} \cdot p_2^{\alpha_2} \dots p_m^{\alpha_m} =$$

$$= \frac{\sqrt{2\pi n} \cdot e^{-n} e^{\theta_n - \sum_{i=1}^{m} \theta_{\alpha_i}} \cdot m}{(2\pi)^{\frac{m}{2}} \prod_{i=1}^{m} \alpha_i^{\alpha_i + \frac{1}{2}} \cdot e^{-\alpha_i}} \prod_{i=1}^{m} p_i^{\alpha_i} = \frac{1}{(2\pi)^{\frac{m-1}{2}}} \sqrt{\frac{n}{\alpha_1!\, \alpha_2! \dots \alpha_m!}}\, e^{\theta} \prod_{i=1}^{m} \left(\frac{np_i}{\alpha_i}\right)^{\alpha_i} =$$

$$= \frac{e^{\theta}}{(2\pi n)^{\frac{m-1}{2}} \sqrt{p_1 \cdot p_2 \dots p_m}} \prod_{i=1}^{m} \left(1 + \xi_i \sqrt{\frac{q_i}{np_i}}\right)^{-\left(np_i + \xi_i \sqrt{npq_i} + \frac{1}{2}\right)}$$

$$\left(\theta = \theta_n - \sum_{i=1}^{m} \theta_{\alpha_i}\right).$$

Pentru evaluarea asimptotică a produsului

$$\prod_{i=1}^{m} \left(1 + \xi_i \sqrt{\frac{q_i}{np_i}}\right)^{-\left(np_i + \xi_i \sqrt{npq_i} + \frac{1}{2}\right)}$$

aplicăm logaritmii

$$\ln \prod_{i=1}^{m} \left(1 + \xi_i \sqrt{\frac{q_i}{np_i}}\right)^{-\left(np_i + \xi_i \sqrt{npq_i} + \frac{1}{2}\right)} =$$

$$= -\sum_{i=1}^{m} \left(np_i + \xi_i \sqrt{npq_i} + \frac{1}{2}\right)\left[\xi_i \sqrt{\frac{q_i}{np_i}} - \frac{1}{2} \frac{q_i \xi_i^2}{np_i} + O\left(n^{-\frac{3}{2}}\right)\right] =$$

$$= -\frac{1}{2} \sum_{i=1}^{m} q_i \xi_i^2 + O\left(n^{-\frac{3}{2}}\right),$$

deoarece

$$\sum_{i=1}^{m} \xi_i \sqrt{np_iq_i} = \sum_{i=1}^{m} (\alpha_i - np_i) = \sum_{i=1}^{m} \alpha_i - n = 0 \qquad \left(\sum_{i=1}^{m} \alpha_i = n\right).$$

In acelaşi timp, ţinînd seama că

$$|\theta| < |\theta_n| + \left|\sum_{i=1}^{m} \theta_{\alpha_i}\right| \leq \frac{1}{12n} + \sum_{i=1}^{n} \frac{1}{12\alpha_i} = \frac{1}{12}\left(\frac{1}{n} + \sum_{i=1}^{m} \frac{1}{\alpha_i}\right) =$$

$$= \frac{1}{12n}\left(1 + \sum_{i=1}^{m} \frac{1}{p_i\left(1 + \xi_i\sqrt{\frac{2}{np_i}}\right)}\right),$$

convergenţa este demonstrată.

§ 4. Momente

Să considerăm şirul de variabile aleatorii independente

$$X_i \begin{pmatrix} a_1, & a_2, \ldots, a_m \\ p_1, & p_2, \ldots, p_m \end{pmatrix} \qquad \begin{matrix} (i = 1, 2, \ldots, n) \\ \left(\sum_{k=1}^{m} p_k = 1\right), \end{matrix}$$

unde $a_1, a_2, \ldots, a_m$ sînt nişte constante date şi să notăm

$$X = \sum_{i=1}^{n} X_i = \sum_{i=1}^{m} a_i\lambda_i \qquad \left(\sum_{i=1}^{m} \lambda_i = n\right).$$

1. *Momentele de ordinul întîi*

$$M(X) = \sum_{i=1}^{n} M(X_i) = n(a_1p_1 + a_2p_2 + \ldots + a_mp_m).$$

In acelaşi timp avem

$$M(X) = a_1M(\lambda_1) + a_2M(\lambda_2) + \ldots + a_mM(\lambda_m).$$

Prin urmare, putem scrie

$$np_1a_1 + np_2a_2 + \ldots + np_ma_m = a_1M(\lambda_1) + a_2M(\lambda_2) + \ldots + a_mM(\lambda_m). \qquad \text{(VIII.4,1)}$$

Egalitatea (VIII.4,1) fiind adevărată oricare ar fi $a_i (i=1, 2, \dots, m)$ rezultă imediat momentele de ordinul întîi

(I) $M(\lambda_1)=np_1,\ M(\lambda_2)=np_2, \dots, M(\lambda_i)=np_i, \dots, M(\lambda_m)=np_m.$

2. *Momentele de ordinul al doilea*

$$M_2(X)=M\left(\sum_{i=1}^{n} X_i\right)^2=\sum_{i=1}^{n} M(X_i^2)+2\sum_{i\neq j} M(X_i)\,M(X_j)=$$

$$=n(a_1^2p_1+a_2^2p_2+\dots+a_m^2p_m)+n(n-1)(a_1p_1+a_2p_2+\dots+a_mp_m)^2=$$

$$=\sum_{i=1}^{m}[np_i+n(n-1)p_i^2]a_i^2+2n(n-1)\sum_{i\neq j} p_ip_ja_ia_j.$$

De asemenea,

$$M_2(X)=M\left(\sum_{i=1}^{m} a_i\lambda_i\right)^2=\sum_{i=1}^{m} a_i^2\,M(\lambda_i^2)+2\sum_{i\neq j} a_ia_j\,M(\lambda_i\lambda_j).$$

Atunci

$$\sum_{i=1}^{m}[np_i+n(n-1)p_i^2]a_i^2+2n(n-1)\sum_{i\neq j} p_ip_ja_ia_j=$$

$$=\sum_{i=1}^{m} M(\lambda_i^2)\,a_i^2+2\sum_{i\neq j} M(\lambda_i\lambda_j)\,a_ia_j. \qquad \text{(VIII.4,2)}$$

Egalitatea (VIII.4,2) avînd loc oricare ar fi $a_i (i=1, 2, \dots, m)$ obţinem, prin identificarea coeficienţilor lui a_i^2 şi a_ia_j, momentele de ordinul al doilea

(II)
$$M(\lambda_i^2)=np_i^2+np_i(1-p_i)$$
$$M(\lambda_i\lambda_j)=n(n-1)p_ip_j. \qquad (i\neq j=1, 2, \dots, m).$$

3. *Momentele de ordinul al treilea*

$$M_3(X)=M\left(\sum_{i=1}^{n} X_i\right)^3=\sum_{i=1}^{n} M(X_i^3)+3\sum_{i\neq j} M(X_i^2)\,M(X_j)+$$

$$+6\sum_{i\neq j\neq k} M(X_i)\,M(X_j)\,M(X_k)=n(a_1^3p_1+\dots+a_m^3p_m)+$$

$$+3n(n-1)(a_1^2p_1+\ldots+a_m^2p_m)(a_1p_1+\ldots+a_mp_m)+$$

$$+n(n-1)(n-2)(a_1p_1+\ldots+a_mp_m)^3=$$

$$=\sum_{i=1}^{m}[np_i+3n(n-1)p_i^2+n(n-1)(n-2)p_i^3]a_i^3+$$

$$+3\sum_{i\neq j}[n(n-1)p_ip_j+n(n-1)(n-2)p_i^2p_j]a_ia_j+$$

$$+6\sum_{i\neq j\neq k}n(n-1)(n-2)p_ip_jp_ka_ia_ja_k.$$

Totodată avem

$$M_3(X)=M\left(\sum_{i=1}^{m}a_i\lambda_i\right)^3=\sum_{i=1}^{m}a_i^3M(\lambda_i^3)+3\sum_{i\neq j}a_i^2a_jM(\lambda_i^2\lambda_j)+$$

$$+6\sum_{i\neq j\neq k}a_ia_ja_kM(\lambda_i\lambda_j\lambda_k).$$

Rezultă, deci,

$$\sum_{i=1}^{m}[np_i+3n(n-1)p_i^2+n(n-1)(n-2)p_i^3]a_i^3+3\sum_{i\neq j}[n(n-1)p_ip_j+$$

$$+n(n-1)(n-2)p_i^2p_j]a_i^2a_j+6\sum_{i\neq j\neq k}n(n-1)(n-2)p_ip_jp_ka_ia_ja_k=$$

$$=\sum_{i=1}^{m}M(\lambda_i^3)a_i^3+3\sum_{i\neq j}M(\lambda_i^2\lambda_j)a_i^2a_j+6\sum_{i\neq j\neq k}M(\lambda_i\lambda_j\lambda_k)a_ia_ja_k.$$

(VIII.4,3)

Din egalitatea (VIII.4,3), care este verificată oricare ar fi a_i $(i=1,2,\ldots,m)$, deducem momentele de ordinul al treilea

$$M(\lambda_i^3)=n^3p_i^3+3n^2p_i^2(1-p_i)+np_i(1-p_i)(1-2p_i),$$

(III)
$$M(\lambda_i^2\lambda_j)=n^3p_i^2p_j+n^2p_ip_j(1-3p_i)-np_ip_j(1-2p_i),$$

$$M(\lambda_i\lambda_j\lambda_k)=n(n-1)(n-2)p_ip_jp_k \quad (i\neq j\neq k=1,\ 2,\ \ldots.\ m).$$

4. *Momentele de ordinul al patrulea*

$$M_4(X)=M\left(\sum_{i=1}^{n}X_i\right)^4=\sum_{i=1}^{n}M(X_i^4)+4\sum_{i\neq j}M(X_i^3)\,M(X_j)+$$

$$+6\sum_{i\neq j}M(X_i^2)\,M(X_j^2)+12\sum_{i\neq j\neq k}M(X_i^2)\,M(X_j)\,M(X_k)+$$

$$+24\sum_{i\neq j\neq k\neq l}M(X_i)\,M(X_j)\,M(X_k)M(X_l)=n(a_1^4p_1+\ldots+a_m^4p_m)+$$

$$+4n(n-1)(a_1^3p_1+\ldots+a_m^3p_m)(a_1p_1+\ldots+a_mp_m)+$$

$$+3n(n-1)(a_1^2p_1+\ldots+a_m^2p_m)^2+6n(n-1)(n-2)(a_1^2p_1+\ldots+a_m^2p_m)\times$$

$$\times(a_1p_1+\ldots+a_mp_m)^2+n(n-1)(n-2)(n-3)(a_1p_1+\ldots+a_mp_m)^4=$$

$$=\sum_{i=1}^{m}[np_i+4n(n-1)p_i^2+3n(n-1)p_i^2+6n(n-1)(n-2)p_i^3+$$

$$+n(n-1)(n-2)(n-3)p_i^4]\,a_i^4+4\sum_{i\neq j}[n(n-1)p_ip_j+3n(n-1)(n-2)\times$$

$$\times(p_i^2p_j+p_ip_j^2)+n(n-1)(n-2)(n-3)p_i^3p_j]\,a_i^3a_j+$$

$$+6\sum_{i\neq j}[n(n-1)p_ip_j+n(n-1)(n-2)(p_i^2p_j+p_ip_j^2)+$$

$$+n(n-1)(n-2)(n-3)p_i^2p_j^2]\,a_i^2a_j^2+12\sum_{i\neq j\neq k}[n(n-1)(n-2)p_ip_jp_k+$$

$$+n(n-1)(n-2)(n-3)p_i^2p_jp_k]\,a_i^2a_ja_k+$$

$$+24\sum_{i\neq j\neq k\neq l}n(n-1)(n-2)(n-3)p_ip_jp_kp_la_ia_ja_ka_l.$$

Momentul de ordinul al patrulea se mai poate scrie

$$M_4(X)=M\left(\sum_{i=1}^{m}a_i\lambda_i\right)^4=\sum_{i=1}^{m}a_i^4M(\lambda_i^4)+4\sum_{i\neq j}a_i^3a_jM(\lambda_i^3\lambda_j)+$$

$$+6\sum_{i\neq j}a_i^2a_j^2M(\lambda_i^2\lambda_j^2)+12\sum_{i\neq j\neq k}a_i^2a_ja_kM(\lambda_i^2)M(\lambda_j)M(\lambda_k)+$$

$$+24\sum_{i\neq j\neq k\neq l}a_ia_ja_ka_lM(\lambda_i)M(\lambda_j)M(\lambda_k)M(\lambda_l).$$

Ținînd seama de cele două expresii ale lui $M_4(X)$, care au loc oricare ar fi $a_i\,(i=1,\ 2,\ \ldots,\ m)$ găsim următoarele momente de ordinul al patrulea, prin identificarea coeficienților lui a_i^4, $a_i^3a_j$, $a_i^2a_j^2$, $a_i^2a_ja_k$ și $a_ia_ja_ka_l$:

$$M(\lambda_i^4)=n^4p_i^4+6n^3p_i^3(1-p_i)+$$

$$+n^2p_i^2(1-p_i)(7-11p_i)-3np_i(1-p_i)(1+2p_i-2p_i^2),$$

$$M(\lambda_i^3\lambda_j)=n^4p_i^3p_j+3n^3p_ip_j(p_i+p_j-2p_i^2)+n^2p_ip_j(1-9p_i-9p_j+$$

$$+11p_i^2)-np_ip_j(1-6p_i-6p_j+6p_i^2),$$

(IV)
$$M(\lambda_i^2\lambda_j^2)=n^4p_i^2p_j^2+n^3p_ip_j(p_i+p_j-6p_ip_j)+$$

$$+n^2p_ip_j(1-3p_i-3p_j+11p_ip_j)-np_ip_j(1-2p_i-2p_j+6p_ip_j),$$

$$M(\lambda_i^2\lambda_j\lambda_k)=n^4p_i^2p_jp_k+n^3p_ip_jp_k(1-6p_i)-n^2p_ip_jp_k(3-11p_i)+$$

$$+2p_ip_jp_k(1-3p_i).$$

$$M(\lambda_i\lambda_j\lambda_k\lambda_l)=n(n-1)(n-2)(n-3)p_ip_jp_kp_l.$$

§ 5. Momente centrate

Ne propunem să determinăm momentele centrate al distribuției multinomiale. Pentru aceasta, considerăm variabilele aleatorii independente X_i din paragraful precedent

$$Y_i=X_i-M(X_i)=X_i-\sum_{i=1}^{m}a_jp_j \qquad (i=1,\ 2,\ \ldots,\ n)$$

și vom nota

$$Y=\sum_{i=1}^{n}Y_i=\sum_{i=1}^{m}\mu_ia_i.$$

1. *Momentele centrate de ordinul întîi*

$$M(Y)=M\left(\sum_{i=1}^{n}Y_i\right)=\sum_{i=1}^{n}M(Y_i)=0.$$

Mai putem scrie

$$M(Y)=M\left(\sum_{i=1}^{m}\mu_i a_i\right)=\sum_{i=1}^{m}M(\mu_i)\,a_i,$$

de unde rezultă că

$$\sum_{i=1}^{m}M(\mu_i)\,a_i=0,$$

oricare ar fi $a_i\,(i=1,\ 2,\ldots,m)$ şi deci

$$M(\mu_i)=0 \qquad (i=1,\ 2,\ldots,m). \tag{I'}$$

2. *Momentele centrate de ordinul al doilea*

$$M_2(Y)=M\left(\sum_{i=1}^{n}Y_i\right)^2=\sum_{i=1}^{n}M(Y_i^2)+2\sum_{i\neq j}M(Y_i)\,M(Y_j)=$$

$$=n\left[M\left(X_i-\sum_{j=1}^{m}a_jp_j\right)^2\right]=n\left[\sum_{i=1}^{m}a_ip_i-\left(\sum_{j=1}^{m}a_jp_j\right)^2\right]=$$

$$=n\sum_{i=1}^{m}p_i(1-p_i)\,a_i^2-2n\sum_{j\neq k}p_jp_ka_ja_k.$$

Pe de altă parte

$$M_2(Y)=M\left(\sum_{i=1}^{m}\mu_ia_i\right)^2=\sum_{i=1}^{m}M(\mu_i^2)\,a_i^2+2\sum_{i\neq j}M(\mu_i\mu_j)\,a_ia_j.$$

Avem deci, oricare ar fi $a_i\,(i=1,\ 2,..,,m)$, egalitatea

$$\sum_{i=1}^{m}np_i(1-p_i)\,a_i^2-2\sum_{i\neq k}np_jp_ka_ja_k=\sum_{i=1}^{m}M(\mu_i^2)\,a_i^2+2\sum_{i\neq j}M(\mu_i\mu_j)\,a_ia_j. \tag{VIII.5,1}$$

Din egalitatea (VIII.5,1) deducem, astfel, momentele centrate de ordinul al doilea

$$\begin{aligned}&M(\mu_i^2)=np_i(1-p_i),\\ &M(\mu_i\mu_j)=-np_ip_j.\end{aligned} \qquad (i\neq j=1,\ 2,\ldots,m). \tag{II'}$$

3. *Momentele centrate de ordinul al treilea*

$$M_3(Y)=M\left(\sum_{i=1}^{n}Y_i\right)^3=\sum_{i=1}^{n}M(Y_i^3)+3\sum_{i\neq j}M(Y_i^2)\,M(Y_j)+$$

$$+6\sum_{i\neq j\neq k}M(Y_i)\,M(Y_j)\,M(Y_k)=nM\left(X_i-\sum_{j=1}^{m}a_jp_j\right)^3=$$

$$=n\left[M(X_i^3)-3\sum_{j=1}^{m}a_jp_jM(X_i^2)+3\left(\sum_{j=1}^{m}a_jp_j\right)^2M(X_i)-\left(\sum_{j=1}^{m}a_jp_j\right)^3\right]=$$

$$=n\left[\sum_{i=1}^{m}a_i^3p_i-3\sum_{j=1}^{m}a_jp_j\sum_{i=1}^{m}a_i^2p_i+2\left(\sum_{j=1}^{m}a_jp_j\right)^3\right]=$$

$$=\sum_{i=1}^{m}np_i(1-3p_i+2p_i^2)\,a_i^3-3\sum_{i\neq j}np_ip_j(1-2p_i)\,a_i^2a_j+$$

$$+12\sum_{i\neq j\neq k}np_ip_jp_ka_ia_ja_k.$$

Insă, momentul centrat de ordinul al treilea se mai poate scrie

$$M_3(Y)=M\left(\sum_{i=1}^{m}\mu_ia_i\right)^3=\sum_{i=1}^{m}M(\mu_i^3)\,a_i^3+3\sum_{i\neq j}M(\mu_i^2\mu_j)+6\sum_{i\neq j\neq k}M(\mu_i\mu_j\mu_k).$$

Egalînd acum cele două expresii ale momentului centrat de ordinul al treilea, care trebuie să fie adevărate pentru orice a_i $(i=1, 2,..., m)$, obținem momentele centrate căutate

$$\text{(III')}\qquad \begin{aligned} &M(\mu_i^3)=np_i(1-p_i)(1-2p_i),\\ &M(\mu_i^2\mu_j)=-np_ip_j(1-2p_i), \qquad (i\neq j=1,\ 2,\ \dots,\ m).\\ &M(\mu_i\mu_j\mu_k)=2np_ip_jp_k. \end{aligned}$$

4. *Momentele centrate de ordinul al patrulea*

$$M_4(Y)=M\left(\sum_{i=1}^{n}Y_i\right)^4=\sum_{i=1}^{n}M(Y_i^4)+4\sum_{i\neq j}M(Y_i^3)\,M(Y_j)+$$

$$+6\sum_{i\neq j}M(Y_i^2)\,M(Y_j^2)+12\sum_{i\neq j\neq k}M(Y_i^2)\,M(Y_j)\,M(Y_k)+$$

$$+24\sum_{i\neq j\neq k\neq l} M(Y_i)\,M(Y_j)\,M(Y_k)\,M(Y_l)=$$

$$=nM\left(X_i-\sum_{j=1}^{m}a_jp_j\right)^4+3n(n-1)\left[M\left(X_i-\sum_{j=1}^{m}a_jp_j\right)^2\right]^2=$$

$$=n\sum_{i=1}^{m}M(X_i^4)-4n\sum_{i=1}^{m}a_jp_jM(X_i^3)+6n\left(\sum_{j=1}^{m}a_jp_j\right)^2M(X_i^2)-$$

$$-4n\left(\sum_{j=1}^{m}a_jp_j\right)^3M(X_i)+n\left(\sum_{j=1}^{m}a_jp_j\right)^4+$$

$$+3n(n-1)\left[\sum_{i=1}^{m}p_i(1-p_i)a_i^2-2\sum_{j\neq k}p_jp_ka_ja_k\right]^2=n\sum_{i=1}^{m}a_i^4p_i-$$

$$-4\sum_{i=1}^{m}a_i^3p_i\sum_{j=1}^{m}a_jp_j+6n\sum_{i=1}^{m}a_i^2p_i\left(\sum_{j=1}^{m}a_j^2p_j^2+2\sum_{j\neq k}a_ja_kp_jp_k\right)-$$

$$-3n\left[\sum_{i=1}^{m}a_i^4p_i^4+4\sum_{i\neq j}a_i^3a_jp_i^3p_j+6\sum_{i\neq j}a_i^2a_j^2p_i^2p_j^2+\right.$$

$$\left.+12\sum_{i\neq j\neq k}a_i^2a_ja_kp_i^2p_jp_k+24\sum_{i\neq j\neq k\neq l}a_ia_ja_ka_lp_ip_jp_kp_l\right]+$$

$$+3n(n-1)\left[\sum_{i=1}^{m}(1-p_i)^2p_i^2a_i^4+2\sum_{i\neq j}(1-p_i)(1-p_j)p_ip_ja_i^2a_j^2-\right.$$

$$\left.-4\sum_{i=1}^{m}(1-p_i)p_ia_i^2\sum_{i\neq k}p_jp_ka_ja_k+4\left(\sum_{i\neq k}p_j^2p_k^2a_j^2a_k^2+2\sum_{i\neq j\neq k\neq l}p_ip_jp_kp_la_ia_ja_ka_l\right)\right].$$

De asemenea avem

$$M_4(Y)=M\left(\sum_{i=1}^{m}\mu_ia_i\right)^4=\sum_{i=1}^{m}M(\mu_i^4)\,a_i^4+4\sum_{i\neq j}M(\mu_i^3\mu_j)\,a_i^3a_j+$$

$$+6\sum_{i\neq j}M(\mu_i^2\mu_j^2)\,a_i^2a_j^2+12\sum_{i\neq j\neq k}M(\mu_i^2\mu_j\mu_k)\,a_i^2a_ja_k+$$

$$+24\sum_{i\neq j\neq k\neq l}M(\mu_i\mu_j\mu_k\mu_l)\,a_ia_ja_ka_l.$$

Din cele două expresii ale lui $M_4(Y)$, deducem egalitatea

$$\sum_{i=1}^{m}[np_i-4np_i^2+6np_i^3-3np_i^4+3n(n-1)(1-p_i)^2p_i^2]a_i^4-$$

$$-4\sum_{i\neq j}[np_ip_j-3np_i^2p_j+3np_i^3p_j+3n(n-1)p_i^2p_j(1-p_i)]a_i^3a_j+$$

$$+6\sum_{i\neq j}[np_ip_j(p_i+p_j)-3np_i^2p_j^2+n(n-1)(1-p_i)(1-p_j)p_ip_j+$$

$$+2n(n-1)p_i^2p_j^2]a_i^2a_j^2+12\sum_{i\neq j\neq k}[np_ip_jp_k-3np_i^2p_jp_k-$$

$$-n(n-1)(1-p_i)p_ip_jp_k]a_ia_ja_k+24\sum_{i\neq j\neq k\neq l}[-3np_ip_jp_kp_l+$$

$$+n(n-1)p_ip_jp_kp_l]a_ia_ja_ka_l=\sum_{i=1}^{m}M(\mu_i^4)a_i^4+4\sum_{i\neq j}M(\mu_i^3\mu_j)a_i^3a_j+$$

$$+6\sum_{i\neq j}M(\mu_i^2\mu_j^2)a_i^2a_i^2+12\sum_{i\neq j\neq k}M(\mu_i^2\mu_j\mu_k)a_i^2a_ja_k+$$

$$+24\sum_{i\neq j\neq k\neq l}M(\mu_i\mu_j\mu_k\mu_l)a_ia_ja_ka_l.$$

Ținînd seama de observațiile făcute la calculul celorlalte momente centrate de ordin inferior, rezultă

$$M(\mu_i^4)=3n^2p_i^2(1-p_i)^2+np_i(1-p_i)(1-6p_i)(1-6p_i+6p_i^2),$$

$$M(\mu_i^3\mu_j)=-3n^2p_i^2p_j(1-p_i)-np_ip_j(1-6p_i+6p_i^2),$$

$$M(\mu_i^2\mu_j^2)=n^2p_ip_j(1-p_i-p_j+3p_ip_j)-np_ip_j(1-2p_i-2p_j+6p_ip_j),$$

$$M(\mu_i^2\mu_j\mu_k)=-n^2p_ip_jp_k(1-p_i)+2np_ip_jp_k(1-2p_i),$$

$$M(\mu_i\mu_j\mu_k\mu_l)=n(n-4)p_ip_jp_kp_l \qquad (i\neq j\neq k\neq l=1,\ 2,\ \ldots,\ m).$$

Folosind același procedeu se pot calcula și celelalte momente centrate de ordin superior ale distribuției multinomiale.

§ 6. Funcţia caracteristică

Să luăm ca bază într-un spaţiu euclidian cu m dimensiuni, vectorii liniar independenţi

$$e_1(1, 0, \ldots, 0, 0),$$
$$e_2(0, 1, \ldots, 0, 0),$$
$$\ldots\ldots\ldots\ldots$$
$$e_m(0, 0, \ldots, 0, 1).$$

Dacă presupunem că

$$\lambda_1, \lambda_2, \ldots, \lambda_m$$

sînt coordonatele unui vector λ în acest spaţiu cu m dimensiuni, vom avea

$$\lambda = \lambda_1 e_1 + \lambda_2 e_2 + \ldots + \lambda_m e_m.$$

Observăm că vectorii $\lambda(\lambda_1, \lambda_2 \ldots, \lambda_m)$ se găsesc într-un subspaţiu liniar cu $m-1$ dimensiuni, deoarece între coordonatele $\lambda_i (i = 1, 2, \ldots, m)$ ale vectorului λ, există relaţia

$$\lambda_1 + \lambda_2 + \ldots + \lambda_m = n.$$

Luînd un sistem de numere reale,

$$u(u_1, u_2, \ldots, u_m),$$

putem defini, în spaţiul euclidian considerat, produsul scalar

$$u\lambda = u_1\lambda_1 + u_2\lambda_2 + \ldots + u_m\lambda_m.$$

Atunci, prin definiţie, *funcţia caracteristică a vectorului* $\lambda(\lambda_1, \ldots, \lambda_m)$ va fi

$$\varphi(u) = M(e^{\lambda u i}) = M[e^{(u_1\lambda_1 + u_2\lambda_2 + \ldots + u_m\lambda_m) i}]. \qquad \text{(VIII.6,1)}$$

Considerăm vectorul $\lambda(\lambda_1, \lambda_2, \ldots, \lambda_m)$ ca o sumă de vectori liniar independenţi

$$X_k \begin{pmatrix} e_1, & e_2, \ldots, e_m \\ p_1, & p_2, \ldots, p_m \end{pmatrix} \qquad \begin{matrix} (k = 1, 2, \ldots, n) \\ \left(\sum_{i=1}^{m} p_i = 1\right). \end{matrix}$$

Funcţia caracteristică a unui vector X_k fiind

$$\psi(u)=p_1e^{u_1i}+p_2e^{u_2i}+\ldots+p_me^{u_mi},$$

rezultă imediat, în baza teoremei privind suma de variabile independente, funcţia caracteristică a vectorului $\lambda=\sum_{k=1}^{n}X_k$

$$\varphi(u)=(p_1e^{u_1i}+p_2e^{u_2i}+\ldots+p_me^{u_mi})^n. \qquad \text{(VIII.6,2)}$$

Din relaţiile (VIII.6,1) şi (VIII.6,2), deducem

$$M\left(e^{\sum_{k=1}^{m}u_k\lambda_ki}\right)=\left(\sum_{k=1}^{m}p_ke^{u_ki}\right)^n. \qquad \text{(VIII.6,3)}$$

Expresia (VIII.6,2) a funcţiei caracteristice se poate determina şi direct în mod simplu.

Intr-adevăr, prin definiţie, avem

$$\varphi(u)=\sum_{\substack{0\\ \lambda_1,\,\lambda_2,\,\ldots,\,\lambda_m}}^{n}P_n(\lambda_1,\lambda_2,\ldots,\lambda_m)\,e^{(u_1\lambda_1+u_2\lambda_2+\ldots+u_m\lambda_m)\,i}=$$

$$=\sum_{\substack{0\\ \lambda_1,\,\lambda_2,\,\ldots,\,\lambda_m}}^{n}\frac{n!}{\lambda_1!\,\lambda_2!\ldots\lambda_m!}p_1^{\lambda_1}\cdot p_2^{\lambda_2}\ldots p_m^{\lambda_m}e^{u_1\lambda_1i}\ldots e^{u_m\lambda_mi}=$$

$$=\sum_{\substack{0\\ \lambda_1,\,\lambda_2,\,\ldots,\,\lambda_m}}^{n}\frac{n!}{\lambda_1!\,\lambda_2!\ldots\lambda_m!}(p_1e^{u_1i})^{\lambda_1}\ldots(p_me^{u_mi})^{\lambda_m}$$

sau

$$\varphi(u)=(p_1e^{u_1i}+p_2e^{u_2i}+\ldots+p_me^{u_mi})^n.$$

Derivînd succesiv în raport cu $u_k\,(k=1,\ 2,\ldots,m)$ funcţia caracteristică $\varphi(u)$ şi făcînd apoi $u_1=u_2=\ldots=u_m=0$, regăsim momentele obişnuite ale distribuţiei multinomiale $M(\lambda_k)$, $M(\lambda_k\lambda_l)$, $M(\lambda_k^3)$ etc.,

$$M(\lambda_k)=\frac{1}{i}\left[\frac{\partial\varphi(u)}{\partial u_k}\right]_{u_1=u_2=\ldots=u_m=0}=$$

$$=[np_k(p_1e^{u_1i}+p_2e^{u_2i}+\ldots+p_me^{u_mi})^{n-1}e^{u_ki}]_{u_1=u_2=\ldots=0}=np_k,$$

$$M(\lambda_k^2)=\frac{1}{i^2}\left[\frac{\partial^2\varphi(u)}{\partial u_k^2}\right]_{u_1=u_2=\ldots=u_m=0}=[n(n-1)p_k^2(p_1e^{u_1 i}+$$

$$+p_2e^{u_2 i}+\ldots+p_me^{u_m i})^{n-2}e^{u_k i}+np_k(p_1e^{u_1 i}+p_2e^{u_2 i}+\ldots$$

$$\ldots+p_me^{u_m i})^{n-1}e^{2u_k i}]_{u_1=u_2=\ldots=u_m=0}=n(n-1)p_k^2+np_k=n^2p_k^2+$$

$$+np_k(1-p_k),$$

$$M(\lambda_k\lambda_l)=\frac{1}{i^2}\left[\frac{\partial^2\varphi(u)}{\partial u_k\partial u_l}\right]_{u_1=u_2=\ldots=u_m=0}=[n(n-1)p_kp_l(p_1e^{u_1 i}+\ldots$$

$$\ldots+p_me^{u_m i})e^{(u_k+u_l)i}]_{u_1=u_2=\ldots=u_m=0}=n(n-1)p_kp_l,$$

. . . . . .

§ 7. Distribuţia limită

Fie

$$\Phi_n(u)=e^{-\sum_{k=1}^{m}\sqrt{np_k}\,u_k i}\left(\sum_{k=1}^{m}p_k e^{\frac{u_k i}{\sqrt{np_k}}}\right)^n \qquad \text{(VIII.7,1)}$$

funcţia caracteristică a vectorului aleator

$$\lambda\left(\frac{\lambda_1-np_1}{\sqrt{np_1}},\ \frac{\lambda_2-np_2}{\sqrt{np_2}},\ \ldots,\ \frac{\lambda_m-np_m}{\sqrt{np_m}}\right).$$

Ne propunem să determinăm legea de distribuţie limită a vectorului aleator pentru $n\to\infty$, adică legea limită a distribuţiei multinomiale, cînd numărul probelor n creşte nemărginit.

Din (VIII.7,1) obţinem

$$\ln\Phi_n(u)=-\sum_{k=1}^{m}\sqrt{np_k}\,u_k i+n\ln\left(\sum_{k=1}^{m}p_k e^{\frac{u_k i}{\sqrt{np_k}}}\right)$$

sau, folosind dezvoltările

$$\ln \Phi_n(u) = -it\sqrt{n}\sum_{k=1}^{m}\sqrt{p_k}\,u_k + n\ln\left[\sum_{k=1}^{m} p_k\left(1+\frac{u_k t}{\sqrt{np_k}}+\frac{u_k^2 t^2}{2!\,np_k}+\right.\right.$$

$$\left.\left.+\theta\frac{u_k^3}{3!(np_k)^{\frac{3}{2}}}\right)\right]_{|\theta|<1} = -it\sqrt{n}\sum_{k=1}^{m}\sqrt{p_k}\,u_k + n\ln\left[1+\frac{t}{\sqrt{n}}\sum_{k=1}^{m}\sqrt{p_k}\,u_k+\right.$$

$$\left.+\frac{t^2}{2n}\sum_{k=1}^{m}u_k^2+O\left(n^{-\frac{3}{2}}\right)\right],$$

unde am notat cu $O\left(n^{-\frac{3}{2}}\right)$ o funcție asimptotic egală cu $n^{-\frac{3}{2}}$. Dezvoltînd în serie

$$\ln\left\{1+\left[\frac{t}{\sqrt{n}}\sum_{k=1}^{m}\sqrt{p_k}u_k+\frac{t^2}{2n}\sum_{k=1}^{m}u_k^2+O\left(n^{-\frac{3}{2}}\right)\right]\right\},$$

avem

$$\ln \Phi_n(u) = -it\sqrt{n}\sum_{k=1}^{m}\sqrt{p_k}\,u_k+\left(it\sqrt{n}\sum_{k=1}^{m}\sqrt{p_k}\,u_k+\frac{t^2}{2}\sum u_k^2\right)-$$

$$-\frac{t^2}{2}\left(\sum_{k=1}^{m}\sqrt{p_k}\,u_k\right)^2+O\left(n^{-\frac{1}{2}}\right).$$

Efectuînd reducerile, rezultă

$$\ln \Phi_n(u) = -\frac{1}{2}\sum_{k=1}^{m}u_k^2+\frac{1}{2}\left(\sum_{k=1}^{m}\sqrt{p_k}\,u_k\right)^2+O\left(n^{-\frac{1}{2}}\right),$$

de unde

$$\lim_{n\to\infty}\Phi_n(u) = e^{-\frac{1}{2}\sum_{k=1}^{m}u_k^2+\frac{1}{2}\left(\sum_{k=1}^{m}\sqrt{p_k}\,u_k\right)^2} = e^{-\frac{1}{2}\left[\sum_{k=1}^{m}u_k^2-\left(\sum_{k=1}^{m}\sqrt{p_k}\,u_k\right)^2\right]}.$$

Scanned with OKEN Scanner

Vom arăta acum că expresia $e^{-\frac{1}{2}\left[\sum_{k=1}^{m} u_k^2-\left(\sum_{k=1}^{m}\sqrt{p_k}u_k\right)^2\right]}$ este funcţia caracteristică a unei legi normale nesingulare cu $m-1$ dimensiuni.
Din inegalitatea

$$\sum_{k=1}^{m} x_k^2 \sum_{k=1}^{m} y_k^2 \geqq \left(\sum_{k=1}^{m} x_k y_k\right)^2$$

deducem, punînd $x_k=\sqrt{p_k}$ şi $y_k=u_k$,

$$\sum_{k=1}^{m} p_k \sum_{k=1}^{m} u_k^2 \geqq \left(\sum_{k=1}^{m}\sqrt{p_k}u_k\right)^2$$

sau

$$\sum_{k=1}^{m} u_k^2-\left(\sum_{k=1}^{m}\sqrt{p_k}u_k\right)^2 \geqq 0. \qquad \text{(VIII.7,2)}$$

Egalitatea are loc, numai dacă

$$\frac{u_1}{\sqrt{p_1}}=\frac{u_2}{\sqrt{p_2}}=\frac{u_3}{\sqrt{p_3}}=\ldots=\frac{u_k}{p_k}=\ldots=\frac{u_m}{\sqrt{p_m}}. \qquad \text{(VIII.7,3)}$$

Relaţia (VIII.7,2) ne arată că expresia $\sum_{k=1}^{m} u_k^2-\left(\sum_{k=1}^{m}\sqrt{p_k}u_k\right)^2$ este o formă definită pozitivă, iar faptul că se anulează numai pentru sistemul de valori diferite de zero, îndeplinind condiţia (VIII.7,3), înseamnă că rangul acestei forme definite pozitiv este $m-1$.
Prin urmare, putem enunţa următoarea
Teoremă. *Vectorul aleator*

$$\lambda\left(\frac{\lambda_1-np_1}{\sqrt{np_1}}, \frac{\lambda_2-np_2}{\sqrt{np_2}}, \ldots, \frac{\lambda_m-np_m}{\sqrt{np_m}}\right)$$

urmează la limită, cînd n tinde către infinit, o lege normală nesingulară cu $m-1$ *dimensiuni.*

§ 8. Funcţia de distribuţie

Funcţia de distribuţie în cazul distribuţiei multinomiale poate căpăta o expresie analitică utilă pentru aplicaţii, în special în teoria selecţiei. M. Stoka a stabilit formula generală, după metoda care urmează. Pentru simplificare, vom considera cazul unei distribuţii multinomiale cu trei dimensiuni. Demonstraţia rămîne aceeaşi în cazul general.

Considerăm funcţia de repartiţie

$$F_{hk}(p,q)=\sum_{\alpha=0}^{h}\sum_{\beta=0}^{k}\frac{n!}{\alpha!\,\beta!\,(n-\alpha-\beta)!}p^{\alpha}q^{\beta}(1-p-q)^{n-\alpha-\beta},$$

unde

$$h<n,\quad k<n,\quad 0\leqslant p\leqslant 1,\quad 0\leqslant q\leqslant 1,\quad 0\leqslant r\leqslant 1,\quad p+q+r=1.$$

Avem

$$\frac{\partial^2 F_{hk}(p,q)}{\partial p\partial q}=\sum_{\alpha=1}^{h}\sum_{\beta=1}^{k}\frac{n!}{(\alpha-1)!\,(\beta-1)!\,(n-\alpha-\beta)!}p^{\alpha-1}q^{\beta-1}(1-p-q)^{n-\alpha-\beta}-$$

$$-\sum_{\alpha=0}^{h}\sum_{\beta=1}^{k}\frac{n!}{\alpha!\,(\beta-1)!\,(n-\alpha-\beta-1)!}p^{\alpha}q^{\beta-1}(1-p-q)^{n-\alpha-\beta-1}- \tag{VIII.8,1}$$

$$-\sum_{\alpha=1}^{h}\sum_{\beta=0}^{k}\frac{n!}{(\alpha-1)!\,\beta!\,(n-\alpha-\beta-1)!}p^{\alpha-1}q^{\beta}(1-p-q)^{n-\alpha-\beta-1}+$$

$$+\sum_{\alpha=0}^{h}\sum_{\beta=0}^{k}\frac{n!}{\alpha!\,\beta!\,(n-\alpha-\beta-2)!}p^{\alpha}q^{\beta}(1-p-q)^{n-\alpha-\beta-2}.$$

Avem relaţiile

$$\sum_{\alpha=1}^{h}\sum_{\beta=1}^{k}\frac{n!}{(\alpha-1)!\,(\beta-1)!\,(n-\alpha-\beta)!}p^{\alpha-1}q^{\beta-1}(1-p-q)^{n-\alpha-\beta}=$$

$$=\sum_{\alpha=0}^{h-1}\sum_{\beta=0}^{k-1}\frac{n!}{\alpha!\,\beta!\,(n-\alpha-\beta-2)!}p^{\alpha}q^{\beta}(1-p-q)^{n-\alpha-\beta-2}=$$

$$=\sum_{\alpha=0}^{h}\sum_{\beta=0}^{k}\frac{n!}{\alpha!\,\beta!\,(n-\alpha-\beta-2)!}p^{\alpha}q^{\beta}(1-p-q)^{n-\alpha-\beta-2}-$$

$$-\sum_{\alpha=0}^{h}\frac{n!}{\alpha!\,k!\,(n-\alpha-k-2)!}p^{\alpha}q^{k}(1-p-q)^{n-\alpha-k-2}-$$

$$-\sum_{\beta=0}^{k}\frac{n!}{h!\,\beta!\,(n-h-\beta-2)!}p^{h}q^{\beta}(1-p-q)^{n-h-\beta-2}+$$

$$+\frac{n!}{h!\,k!\,(n-h-k-2)!}p^h q^k (1-p-q)^{n-h-k-2}\times$$

$$\times\sum_{\alpha=0}^{h}\sum_{\beta=1}^{k}\frac{n!}{\alpha!\,(\beta-1)!\,(n-\alpha-\beta-1)!}p^\alpha q^{\beta-1}(1-p-q)^{n-\alpha-\beta-1}=$$

$$=\sum_{\alpha=0}^{h}\sum_{\beta=0}^{k-1}\frac{n!}{\alpha!\,\beta!\,(n-\alpha-\beta-2)!}p^\alpha q^k (1-p-q)^{n-\alpha-\beta-2}=$$

$$=\sum_{\alpha=0}^{h}\sum_{\beta=0}^{k}\frac{n!}{\alpha!\,\beta!\,(n-\alpha-\beta-2)!}p^\alpha q^\beta (1-p-q)^{n-\alpha-\beta-2}-$$

$$-\sum_{\alpha=0}^{h}\frac{n!}{\alpha!\,k!\,(n-\alpha-k-2)!}p^\alpha q^k (1-p-q)^{n-\alpha-k-2};$$

$$\sum_{\alpha=1}^{h}\sum_{\beta=0}^{k}\frac{n!}{(\alpha-1)!\,\beta!\,(n-\alpha-\beta-1)!}p^{\alpha-1}q^\beta (1-p-q)^{n-\alpha-\beta-1}=$$

$$=\sum_{\alpha=0}^{h-1}\sum_{\beta=0}^{k}\frac{n!}{\alpha!\,\beta!\,(n-\alpha-\beta-2)!}p^\alpha q^\beta (1-p-q)^{n-h-\beta-2}=$$

$$=\sum_{\alpha=0}^{h}\sum_{\beta=0}^{k}\frac{n!}{\alpha!\,\beta!\,(n-\alpha-\beta-2)!}p^\alpha q^\beta (1-p-q)^{n-\alpha-\beta-2}-$$

$$-\sum_{\beta=0}^{k}\frac{n!}{h!\,\beta!\,(n-h-\beta-2)!}p^h q^\beta (1-p-q)^{n-h-\beta-2}.$$

Inlocuind aceste relații în (VIII.8,1) obținem

$$\frac{\partial^2 F_{hk}(p,q)}{\partial p\,\partial q}=\frac{n!}{h!\,k!\,(n-h-k-2)!}p^h q^k (1-p-q)^{n-h-k-2}.$$

Deci

$$F_{hk}(p,q)=\frac{n!}{h!\,k!\,(n-h-k-2)!}\int_0^p\int_0^q x^h y^k (1-x-y)^{n-h-k-2}\,dxdy+$$

$$+f(p)+\varphi(q)+C. \qquad \text{(VIII.8,2)}$$

Avem

$$F_{hk}(0,0)=1,$$

$$F_{hk}(0,q)=\sum_{\beta=0}^{k}\frac{n!}{\beta!\,(n-\beta)!}\,q^{\beta}(1-p)^{n-\beta}, \qquad \text{(VIII.8,3)}$$

$$F_{hk}(p,0)=\sum_{\alpha=0}^{k}\frac{n!}{\alpha!\,(n-\alpha)!}\,p^{\alpha}(1-p)^{n-\alpha}.$$

Notînd integrala din formula (VIII.8,2) cu $I_{hk}(p,q)$, avem

$$I_{hk}(0,0)=I_{hk}(0,q)=I_{hk}(p,0)=0. \qquad \text{(VIII.8,4)}$$

Avînd în vedere (VIII.8,3) și (VIII.8,4) rezultă

$$C=1-f(0)-\varphi(0),$$

deci

$$F_{hk}(p,q)=I_{hk}(p,q)+1+f(p)-f(0)+\varphi(q)-\varphi(0).$$

Din (VIII.8,3) și (VIII.8,4) rezultă de asemenea

$$f(p)-f(0)=\sum_{\alpha=0}^{h}\frac{n!}{\alpha!\,(n-\alpha)!}\,p^{\alpha}(1-p)^{n-\alpha}-1,$$

$$\varphi(q)-\varphi(0)=\sum_{\beta=0}^{k}\frac{n!}{\beta!\,(n-\beta)!}\,q^{\beta}(1-q)^{n-\beta}-1.$$

Deci

$$F_{hk}(p,q)=I_{hk}(p,q)+\sum_{\alpha=0}^{h}\frac{n!}{\alpha!\,(n-\alpha)!}\,p^{\alpha}(1-p)^{n-\alpha}+$$

$$+\sum_{\beta=0}^{k}\frac{n!}{\beta!\,(n-\beta)!}\,q^{\beta}(1-q)^{n-\beta}-1. \qquad \text{(VIII.8,5)}$$

Se poate arăta, derivînd în raport cu p în ambii membri și apoi prin identificare, că avem

$$\sum_{\alpha=0}^{h}\frac{n!}{\alpha!\,(n-\alpha)!}\,p^{\alpha}(1-p)^{n-\alpha}=1-\frac{n!}{h!\,(n-h-1)!}\int_{0}^{p}x^{h}(1-x)^{n-h-1}\,dx.$$

Inlocuind această valoare în (VIII.8,5), obținem

$$F_{hk}(p,q)=I_{hk}(p,q)-\frac{n!}{h!\,(n-h-1)!}\int_0^p x^h(1-x)^{n-h-1}\,dx-$$

$$-\frac{n!}{k!\,(n-k-1)!}\int_0^q x^k(1-x)^{n-k-1}\,dx+1.$$

Deci

$$\sum_{\alpha=0}^{h}\sum_{\beta=0}^{k}\frac{n!}{\alpha!\,\beta!\,\gamma!}p^\alpha q^\beta r^\gamma=1-\frac{n!}{h!\,(n-h-1)!}\int_0^p x^h(1-x)^{n-h-1}\,dx-$$

$$-\frac{n!}{k!\,(n-k-1)!}\int_0^q x^k(1-x)^{n-k-1}\,dx+$$

$$+\frac{n!}{h!\,k!\,(n-h-k-2)!}\int_0^p\int_0^q x^h y^k(1-x-y)^{n-h-k-2}\,dx\,dy,$$

$$\alpha+\beta+\gamma=n,$$
$$p+q+r=1.$$

In mod analog se demonstrează formula generală

$$\sum_{\alpha_1=0}^{h_1}\sum_{\alpha_2=0}^{h_2}\cdots\sum_{\alpha_m=0}^{h_m}\frac{n!}{\alpha_1!\,\alpha_2!\ldots\alpha_m!\,\alpha_{m+1}!}p_1^{\alpha_1}p_2^{\alpha_2}\ldots p_m^{\alpha_m}p_{m+1}^{\alpha_{m+1}}=$$

$$=1-\sum_{i=1}^{m}\frac{n!}{h_i!\,(n-h_i-1)!}\int_0^{p_i}x^{h_i}(1-x)^{n-h_i-1}dx+$$

$$\sum_{i_1<i_2}^{1,m}\frac{n!}{h_{i_1}!\,h_{i_2}!\,(n-h_{i_1}-h_{i_2}-2)!}\int_0^{p_{i_1}}\int_0^{p_{i_2}}x_1^{h_{i_1}}x_2^{h_{i_2}}(1-x_1-x_2)^{n-h_{i_1}-h_{i_2}-2}\,dx_1\,dx_2-\cdots$$

$$\ldots\pm(-1)^\mu\sum_{i_1<i_2<\ldots<i_\mu}^{1,m}\frac{n!}{h_{i_1}!\,h_{i_2}!\ldots h_{i_\mu}!\,(n-h_{i_1}-h_{i_2}-\ldots-h_{i_\mu}-\mu)!}\times$$

$$\times\int_0^{p_{i_1}}\ldots\int_0^{p_{i_\mu}}x_1^{h_{i_1}}x_2^{h_{i_2}}\ldots x_\mu^{h_{i_\mu}}(1-x_1-x_2-\ldots-x_\mu)^{n-h_{i_1}-h_{i_2}-\ldots-h_{i_\mu}-\mu}\times$$

$$\times dx_1\, dx_2 \ldots dx_\mu \pm \ldots \pm (-1)^m \frac{n!}{h_1!\, h_2! \ldots h_m!\,(n-h_1-h_2-\ldots-h_m-m)!} \times$$

$$\times \int_0^{p_1} \ldots \int_0^{p_m} x_1^{h_1} x_2^{h_2} \ldots x_m^{h_m} (1-x_1-x_2-\ldots-x_m)^{n-h_1-h_2-\ldots-h_m-m} \times$$

$$\times dx_1\, dx_2 \ldots dx_m,$$

$$\alpha_1+\alpha_2+\ldots+\alpha_{m+1}=n,$$

$$p_1+p_2+\ldots+p_{m+1}=1.$$

Capitolul IX

Distribuția hipergeometrică

§ 1. Considerații generale

In aplicațiile statisticii matematice se ivesc probleme care nu pot fi încadrate în schema lui J. Bernoulli şi deci, rezolvarea lor cere în mod necesar cunoaşterea altor tipuri de scheme probabilistice.

Una dintre aceste scheme este *distribuția hipergeometrică*.

Să presupunem că vrem să stabilim, în general, frecvența unei caracteristici pentru o colectivitate de N elemente şi că nu avem posibilitatea să constatăm existența sau lipsa acestei caracteristici, decît pentru un număr $n\,(n<N)$ de aceleaşi elemente. Pornind de la acest rezultat parțial, se naşte întrebarea: ce concluzii putem trage pentru întreaga colectivitate formată din cele N elemente?

Această problemă duce de fapt la schema probabilistică a selecției, de care ne ocupăm pe larg într-unul din capitolele următoare.

Aici construim doar următoarea schemă generală, care poate să servească drept model pentru fenomenele statistice de felul celor arătate mai sus.

Să presupunem o urnă, care conține A bile albe şi B bile negre $(A+B=N)$. Din această urnă se fac n extracții succesive, *fără să se pună bila extrasă înapoi în urnă*. Obținem, în modul acesta, n probe dependente, adică un şir de variabile vectoriale aleatorii de-

pendente, unde fiecare dintre ele poate lua valorile $\xi_{k_1}(1,0)$ sau $\xi_{k_2}(0,1)$ cu probabilitățile respective p_{k_1} și p_{k_2} $(k=1,2,\ldots,n)$, $(p_{k_1}+p_{k_2}=1)$, care urmează să fie determinate.

Mai general, putem să ne închipuim o urnă care conține N_1 bile de culoarea întîi, N_2 bile de culoarea a doua, N_3 bile de culoarea a treia etc. $(N_1+N_2+N_3+\ldots+N_m=N)$. Facem n extracții succesive, fără a pune bila extrasă înapoi în urnă.

In succesiunea de variabile vectoriale aleatorii, astfel obținută, fiecare corespunde unei distribuții

$$X_k\begin{pmatrix}\xi_{k_1}(1,0,\ldots,0), & \xi_{k_2}(0,1,0,\ldots,0), & \ldots, & \xi_m(0,0,\ldots,0,1)\\ p_{k_1} & p_{k_2} & \ldots, & p_{k_m}\end{pmatrix}$$

$$(k=1,2,\ldots,n);\quad \left(\sum_{i=1}^{m} p_{k_i}=1\right).$$

Variabilele X_k sînt de data aceasta dependente, spre deosebire de variabilele similare întîlnite în cazul urnei lui Bernoulli. Natura dependenței acestor variabile este ușor de stabilit.

După ce am făcut k extracții succesive, au mai rămas în urnă $N-k$ bile. Rezultă că avem în urnă $(N-k)p_i$ bile de culoare i $(i=1,2,\ldots,m)$. Să presupunem că extracția de ordinul k ne-a dat o bilă de culoarea h. In urnă au rămas $N-k-1$ bile repartizate în modul următor: $(N-k)p_i$ bile de culoarea i, pentru i luînd toate valorile $1,2,\ldots,m$ cu excepția $i=h$, în care caz avem în urnă $(N-n)p_h-1$ bile de culoare h. Prin urmare, probabilitățile $p_{(k+1)_i}$ de a obține în extracția de rangul $k+1$ o bilă de culoarea i sînt date de relațiile

$$p_{(k+1)_i}=\frac{(N-n)p_i}{N-n-1},\qquad (i=1,2,\ldots,m;\ i\neq h).$$

$$p_{(k+1)_h}=\frac{(N-n)p_h-1}{N-n-1}.$$

Probabilitățile variabilei X_{k+1} depind numai de probabilitățile variabilei X_k și de valoarea luată de X_k. Sîntem deci în prezența unui lanț cu legături complete de ordinul întîi.

§ 2. Determinarea probabilității ca în n probe succesive să obținem α bile albe și β bile negre

Într-o urnă se găsesc A bile albe și B bile negre $(A+B=N)$. Din urnă se fac n extracții succesive, fără să se pună bila extrasă înapoi în urnă.

Teoremă. *Probabilitatea ca în n extracții succesive să scoatem α bile albe și β bile negre $(\alpha+\beta=n)$ este*

$$P_n(\alpha)=\frac{C_A^{\alpha}\cdot C_B^{\beta}}{C_{A+B}^{\alpha+\beta}}.$$

Pentru demonstrarea teoremei, va trebui să determinăm pe de o parte numărul cazurilor posibile, iar pe de altă parte, numărul cazurilor favorabile.

Numărul cazurilor posibile este, evident, numărul combinărilor ce se pot face cu numărul total de bile din urnă, luate în raport cu numărul bilelor extrase, adică

$$C_N^n=C_{A+B}^{\alpha+\beta}.$$

Numărul cazurilor favorabile va fi dat de numărul grupelor ce vom putea forma de câte α bile albe și β bile negre.

Din cele A bile albe, vom putea forma C_A^{α} grupe, care asociate cu C_B^{β} grupe din B bile negre, ne dau $C_A^{\alpha}\cdot C_B^{\beta}$ grupe de câte α bile albe și β bile negre.

Atunci, probabilitatea căutată va fi

$$P_n(\alpha)=\frac{C_A^{\alpha}\cdot C_B^{\beta}}{C_{A+B}^{\alpha+\beta}}. \qquad \text{(IX.2,1)}$$

Teorema este, prin urmare, demonstrată.

1) De remarcat că

$$\sum_{\alpha=a}^{b} P_n(\alpha)=1.$$

Trebuie dată o deosebită atenție limitelor inferioară și superioară de variație a lui α. Limita inferioară a va fi $\max(0,\ n-B)$, iar cea superioară b va fi $\min(n,\ A)$.

2) Pentru $\alpha=0$ avem

$$P_n(0)=\frac{B(B-1)\ldots(B-n+1)}{(A+B)(A+B-1)\ldots(A+B-n+1)}.$$

Folosind această relație, expresia (IX.2,1) se poate pune sub forma

$$P_n(\alpha)=\frac{n(n-1)\dots(n-\alpha+1)}{\alpha!}\cdot\frac{A(A-1)\dots(A-\alpha+1)}{(B-n+1)(B-n+2)\dots(B-\beta)}\cdot P_n(0). \qquad \text{(IX.2,2)}$$

Considerînd, acum, funcția generatoare

$$\Phi(t)=\sum_{\alpha=0}^{n}P_n(\alpha)\,t^{\alpha}$$

și notînd $n=-a$, $A=b$, $B-n+1=c$, rezultă, ținînd seama de (IX.2,2),

$$\Phi(t)=\left[1+\frac{ab}{1\cdot c}t+\frac{a(a+1)\,b(b+1)}{1\cdot 2\cdot c^2}t^2+\dots\right]P_n(0).$$

Expresia din paranteză este seria hipergeometrică a lui Gauss de la care, de altfel, vine și numele distribuției hipergeometrice.

3. Dezvoltînd expresia (IX.2,1) a probabilității $P_n(\alpha)$, obținem

$$P_n(\alpha)=\frac{A(A-1)\dots(A-\alpha+1)}{\alpha!}\cdot\frac{B(B-1)\dots(B-\beta+1)}{\beta!}\cdot\frac{n!}{N(N-1)\dots(N-n+1)}.$$

Sau însemnînd $p=\frac{A}{N}$ și $q=\frac{B}{N}$ deducem

$$P_n(\alpha)=\frac{Np(Np-1)\dots(Np-\alpha+1)\cdot Nq(Nq-1)\dots(Nq-\beta+1)}{N(N-1)\dots(N-n+1)}\cdot\frac{n!}{\alpha!\,\beta!},$$

de unde,

$$\lim_{n\to\infty}P_n(\alpha)=\lim_{N\to\infty}\frac{Np(Np-1)\dots(Np-\alpha+1)\cdot Nq(Nq-1)\dots(Nq-\beta+1)}{N(N-1)\dots(N-n+1)}\cdot\frac{n!}{\alpha!\,\beta!}=$$

$$=\frac{n!}{\alpha!\,\beta!}p^{\alpha}q^{\alpha}.$$

Cu alte cuvinte, cînd numărul total al bilelor N din urnă crește nemărginit, valoarea probabilității $P_n(\alpha)$ după schema bilei neîntoarse tinde către valoarea aceleiași probabilități după schema urnei lui J. Bernoulli.

Acest rezultat era de așteptat, deoarece cu cît numărul inițial de bile din urnă este mai mare, cu atît scoaterea unei bile din urnă exercită o acțiune mai mică asupra compoziției urnei, la limită ajungîndu-se la o succesiune de variabile independente.

Scanned with OKEN Scanner

§. 3. Momente

Să considerăm o urnă care conţine A bile albe şi B bile negre $(A+B=N)$. Din această urnă facem n extracţii succesive, fără să punem bila extrasă înapoi în urnă.

Avem deci, o succesiune de variabile aleatorii dependente

$$X_1,\ X_2,\ \ldots,\ X_i,\ \ldots,\ X_n,$$

fiecare dintre ele luînd valorile $\xi_{i_1}(1,\ 0)$ şi $\xi_{i_2}(0,\ 1)$, cu probabilităţile respective

$$p_{i_1} \quad \text{şi} \quad p_{i_2} \quad (p_{i_1}+p_{i_2}=1), \qquad (i=1,\ 2,\ \ldots,\ n).$$

Intr-o singură probă, variabila vectorială X_i poate să apară sau nu, de unde rezultă că vom avea distribuţia

$$X_i\begin{pmatrix} 1 & 0 \\ p_{i_1} & p_{i_2} \end{pmatrix}.$$

Ne propunem să calculăm primele momente ale variabilei vectoriale sumă

$$X=X_1+X_2+\ldots+X_n.$$

1. *Momentele de ordinul întîi*

$$M_1=M(X)=\sum_{i=1}^{n} M(X_i)=n\sum_{i=1}^{N} p_i. \qquad \text{(IX.3,1)}$$

Pentru determinarea probabilităţii $P(X_i=1)$ ca în extracţia de rangul i să iasă o bilă albă, vom raţiona în felul următor.

Să presupunem că am numerotat cele N bile din urnă de la 1 pînă la N şi am aranjat toate aceste bile în ordine

$$\underbrace{1,\ 2,\ \ldots,\ A-1,\ A}_{\text{bile albe}},\quad \underbrace{A+1,\ \ldots,\ N}_{\text{bile negre}}$$

şi că rezultatele extracţiilor succesive s-au prezentat astfel

$$\xi_1,\ \xi_2,\ \ldots,\ \xi_i,\ \ldots,\ \xi_j,\ \ldots,\ \xi_N,$$

unde numerele $\xi_i\ (i=1,\ 2,\ \ldots,\ N)$ reprezintă numerele $1,\ 2,\ \ldots,\ N$ luate cîte o singură dată.

Vom număra întîi cazurile posibile. Numărul lor va fi egal cu $N!$, adică cu numărul permutărilor celor N bile, deoarece fiecare dintre aceste bile poate să apară în extracţia de rangul i ($i=1, 2, \ldots, N$). Acum trecem la numărarea cazurilor favorabile.

O singură bilă albă apărută la extracţia de rangul i se poate asocia cu oricare dintre grupele de $N-1$ bile, apărute în extracţiile de celelalte ranguri diferite de i. Numărul acestor grupe va fi $(N-1)!$. Cum fiecare dintre cele A bile albe, care pot ieşi în extracţia de rangul i, se poate asocia la rîndul ei cu cele $(N-1)!$ grupe, rezultă că numărul cazurilor favorabile va fi egal cu $A(N-1)!$.

Prin urmare, probabilitatea căutată este

$$P(X_i=1)=\frac{A(N-1)!}{N!}=\frac{A}{N}=p.$$

Se observă că această probabilitate este independentă de i.

Atunci, expresia (IX.3,1) care dă momentul de ordinul întîi se va scrie

$$M_1=np.$$

2. *Momentele de ordinul al doilea*

$$M_2=M(X^2)=M\left(\sum_{i=1}^{n}X_i\right)^2=\sum_{i=1}^{n}M(X_i^2)+2\sum_{i\neq j}M(X_iX_j)=np+$$

$$+n(n-1)M(X_iX_j). \qquad \text{(IX.3,2)}$$

Pentru calculul lui $M(X_iX_j)$ va trebui să considerăm variabila vectorială X_iX_j, care poate lua valorile

$$(X_i=1,\ X_j=1),\ (X_i=1,\ X_j=0),\ (X_i=0,\ X_j=1),\ (X_i=0,\ X_j=0).$$

De observat că produsul X_iX_j ia singura valoare 1 diferită de zero cu probabilitatea $P(X_i=1,\ X_j=1)$, pe care urmează s-o determinăm.

Numărul cazurilor posibile va fi evident $N!$, deoarece în fiecare din extracţiile de rangul i şi j poate să iasă oricare dintre cele N bile din urnă.

Numărul cazurilor favorabile se va obţine asociind în toate modurile posibile cele două bile albe ieşite în extracţiile de rangul i şi j cu oricare dintre cele $(N-2)!$ grupe ce se pot forma cu restul de $N-2$

Scanned with OKEN Scanner

bile apărute în celelalte extracții. In extracțiile de rangul i și j, putînd apare oricare dintre cele A bile albe, înseamnă că numărul cazurilor favorabile va fi egal cu $A(A-1)(N-2)!$

Vom avea astfel

$$P(X_i=1,\ X_j=1)=\frac{A(A-1)(N-2)!}{N!}=\frac{A(A-1)}{N(N-1)}=\frac{p(Np-1)}{N-1}.$$

După cum se vede, această probabilitate nu depinde de i și j.

Ținînd seama de valoarea acestei probabilități, relația (IX.3,2) devine

$$M_2=np+n(n-1)\frac{(Np-1)p}{N-1}=\frac{Np-1}{N-1}pn^2+\frac{N}{N-1}pqn=$$

$$=[(Np-1)pn^2+Npqn]\frac{1}{N-1},$$

unde am notat $q=1-p$.

3. *Momentele de ordinul al treilea*

$$M_3=M(X^3)=M\left(\sum_{i=1}^{n}X_i\right)^3=\sum_{i=1}^{n}M(X_i^3)+3\sum_{i\neq j}M(X_i^2X_j)+$$

$$+6\sum_{i\neq j\neq k}M(X_iX_jX_k)=np+3n(n-1)M(X_i^2X_j)+$$

$$+n(n-1)(n-2)M(X_iX_jX_k). \qquad \text{(IX.3,3)}$$

Urmează, deci, să calculăm pe $M(X_i^2X_j)$ și $M(X_iX_jX_k)$.

In acest caz, va trebui să considerăm variabilele vectoriale $X_i^2X_j$ și $X_iX_jX_k$.

Prima dintre ele poate lua valorile

$$(X_i^2=1,\ X_j=1),\ (X_i^2=1,\ X_j=0),\ (X_i^2=0,\ X_j=1),\ (X_i^2=0,\ X_j=0),$$

iar a doua, valorile următoare

$$(X_i=1,\ X_j=1,\ X_k=1),$$

$$(X_i=1,\ X_j=1,\ X_k=0),\ldots,(X_i=0,\ X_j=0,\ X_k=0).$$

Scanned with OKEN Scanner

După cum se observă, fiecare dintre produsele $X_i^2 X_j$ şi $X_i X_j X_k$ iau numai o singură dată o valoare diferită de zero şi anume valoarea 1.

Rămîne, deci, să determinăm probabilităţile $P(X_i^2=1,\ X_j=1)$ şi $P(X_i=1,\ X_j=1,\ X_k=1)$.

Probabilitatea $P(X_i^2=1,\ X_j=1)$ se deduce imediat. Ea are ca valoare

$$\frac{A(A-1)}{N(N-1)} = \frac{p(Np-1)}{N-1}.$$

Un raţionament identic cu cele anterioare ne va conduce la determinarea celeilalte probabilităţi $P(X_i=1,\ X_j=1,\ X_k=1)$, a cărei valoare o vom găsi egală cu

$$\frac{A(A-1)(A-2)}{N(N-1)(N-2)} = \frac{p(Np-1)(Np-2)}{(N-1)(N-2)}$$

independentă de i, j şi k.

Inlocuind acum în (IX.3,3) pe $M(X_i^2 X_j)$ şi $M(X_i X_j X_k)$ cu valorile lor, care se pot calcula uşor, cunoscînd probabilităţile

$P(X_i^2=1,\ X_j=1)$ şi $P(X_i=1,\ X_j=1,\ X_k=1)$ obţinem

$$M_3 = np + 3n(n-1)\frac{(Np-1)p}{N-1} + n(n-1)(n-2)\frac{(Np-1)(Np-2)p}{(N-1)(N-2)}.$$

Sau, efectuînd calculele şi ordonînd polinomul din membrul al doilea după puterile descrescătoare ale lui n

$$M_3 = [(Np-1)(Np-2)pn^3 + 3N(Np-1)pqn^2 + N^2(1-2p)pqn]\frac{1}{(N-1)(N-2)}.$$

Folosind acelaşi procedeu, putem calcula şi celelalte momente de ordin superior lui trei.

Constatăm că, la limită, pentru $N\to\infty$, momentele M_2 şi M_3 de mai sus tind către momentele obişnuite ale distribuţiei binomiale.

In adevăr,

$$\lim_{N\to\infty} M_2 = \lim_{N\to\infty} [(Np-1)pn^2 + Npqn]\frac{1}{N-1} = p^2n^2 + pqn.$$

$$\lim_{N\to\infty} M_3 = \lim_{N\to\infty} [(Np-1)(Np-2)pn^3+3N(Np-1)pqn^2+$$
$$+N^2(1-2p)pqn]\frac{1}{(N-1)(N-2)} = p^3n^3+3p^2qn^2+(1-2p)pqn.$$

Asemenea rezultate le putem explica ușor, deoarece așa cum am mai arătat și în cazul probabilității $P_n(\alpha)$, cînd numărul bilelor N din urnă crește la infinit, compoziția urnei rămîne aceeași după fiecare extracție. În acest caz, avem schema probabilistică a distribuției binomiale obișnuite, cu probabilități constante în fiecare probă.

§ 4. Momente centrate

Se dă în acest caz succesiunea de variabile aleatorii dependente

$$Y_i\begin{pmatrix}1-p & -p\\ p_{i_1} & p_{i_2}\end{pmatrix} \qquad (i=1,\ 2,\ \ldots,\ n)$$

și se cere să calculăm momentele variabilei vectoriale aleatorii

$$Y=Y_1+Y_2+\ldots+Y_n.$$

1. *Momentul centrat de ordinul întîi*

$$m_1=M(Y)=\sum_{i=1}^{n} M(Y_i)=0.$$

2. *Momentul centrat de ordinul al doilea*

$$m_2=M(Y^2)=M\left(\sum_{i=1}^{n} Y_i\right)^2=\sum_{i=1}^{n} M(Y_i^2)+2\sum_{i\neq j} M(Y_iY_j)=npq+$$
$$+n(n-1)M(Y_iY_j). \qquad \text{(IX.4,1)}$$

Calculul lui $M(Y_iY_j)$ se face ușor. Variabila vectorială Y_iY_j dă naștere la distribuția

$$Y_iY_j=\begin{pmatrix}(Y_i=q,\ Y_j=q), & (Y_i=q,\ Y_j=-p),\\ \frac{(Np-1)p}{N-1}, & \frac{Npq}{N-1},\\ (Y_i=-p,\ Y_j=q,) & (Y_i=-p,\ Y_j=-p)\\ \frac{Npq}{N-1}, & \frac{(Nq-1)q}{N-1}\end{pmatrix}.$$

Atunci, vom avea

$$M(Y_iY_j)=\frac{(Np-1)\,p}{N-1}q^2-2\frac{Npq}{N-1}pq+\frac{(Nq-1)\,q}{N-1}p^2=-\frac{pq}{N-1}\cdot$$

Introducînd în (IX.4,1) valoarea lui $M(Y_iY_j)$, obţinem

$$m_2=npq-n(n-1)\frac{pq}{N-1}=npq\left(1-\frac{n-1}{N-1}\right)=\frac{N-n}{N-1}\cdot npq.$$

Se observă că dispersia corespunzătoare la n probe din distribuţia hipergeometrică diferă de dispersia de la urna lui Bernoulli prin factorul $\frac{N-n}{N-1}$, care este întotdeauna mai mic decît unu. Dacă $N=n$, m_2 devine egal cu zero, adică elementul aleator dispare.

In mod analog, putem obţine şi celelalte momente centrate de ordin superior lui doi.

Observăm că

$$\lim_{N\to\infty} m_2=\lim_{N\to\infty}\frac{N-n}{N-1}npq=npq,$$

adică, la limită, pentru $N\to\infty$, momentul centrat m_2 al distribuţiei hipergeometrice tinde către momentul centrat de ordinul al doilea al distribuţiei binomiale.

Acest rezultat este firesc pentru aceleaşi consideraţii pe care le-am făcut şi mai înainte.

§ 5. Momente factoriale

Am arătat (cap. II, § 6) că momentele factoriale ale unei distribuţii pot fi obţinute din relaţia

$$M_{(h)}=\frac{d^h}{dt^h}[\Phi(t)]_{t=1}\qquad(h=1,\ 2,\ \ldots),$$

unde $\Phi(t)$ este funcţia generatoare

$$\int_{-\infty}^{+\infty}t^{\alpha}dF(\alpha).$$

In cazul distribuţiei hipergeometrice, care este o distribuţie discretă, funcţia generatoare se va scrie

$$\Phi_n(t)=\sum_{a=0}^{n}\frac{C_A^a\cdot C_B^{n-a}}{C_N^n}\,t^a \qquad (A+B=N). \tag{IX.5,1}$$

Să considerăm, acum, expresia

$$(1+tx)^A(1+x)^B.$$

Ţinînd seama de (IX.5,1) observăm că avem dezvoltarea

$$(1+tx)^A(1+x)^B=1+C_N^1\Phi_1(t)\,x+\ldots+C_N^n\Phi_n(t)\,x^n+\ldots$$
$$\ldots+C_N^N\Phi_n(t)x^N. \tag{IX.5,2}$$

Derivînd relaţia (IX.5,2) succesiv de h ori în raport cu t, obţinem

$$A(A-1)\ldots(A-h+1)\,(1+tx)^{A-h}\,x^h(1+x)^B=\sum_{n=1}^{N}C_N^n\frac{d^h}{dt^h}[\Phi_n(t)]\,x^n. \tag{IX.5,3}$$

Dacă facem $t=1$, relaţia (IX.5,3) devine

$$A(A-1)\ldots(A-h+1)\,x^h(1+x)^{N-h}=\sum_{n=1}^{N}C_N^n\frac{d^h}{dt^h}[\Phi_n(t)]_{t=1}\,x^n.$$

Identificînd în ultima egalitate coeficienţii lui x^n din ambii membri, deducem

$$A(A-1)\ldots(A-h+1)\,C_{N-h}^{n-h}=C_N^n\frac{d^h}{dt^h}[\Phi_n(t)]_{t=1},$$

de unde rezultă

$$M_{(h)}=\left[\frac{d^h}{dt^h}\Phi_n(t)\right]_{t=1}=\frac{A(A-1)\ldots(A-h+1)\,C_{N-h}^{n-h}}{C_N^n}$$

sau, efectuînd toate simplificările

$$M_{(h)}=\frac{A(A-1)\ldots(A-h+1)\,n(n-1)\ldots(n-h+1)}{N(N-1)\ldots(N-h+1)}=$$
$$=\frac{p(Np-1)\ldots(Np-h+1)\,n(n-1)\ldots(n-h+1)}{(N-1)\ldots(N-h+1)}.$$

Dînd lui h valorile 1, 2, 3, ..., avem momentele factoriale de diferite ordine.

Astfel, primele momente factoriale sînt

$$M_{(1)}=np,$$

$$M_{(2)}=n(n-1)\frac{p(Np-1)}{N-1},$$

$$M_{(3)}=n(n-1)(n-2)\frac{p(Np-1)(Np-2)}{(N-1)(N-2)},$$

$$M_{(4)}=n(n-1)(n-2)(n-3)\frac{p(Np-1)(Np-2)(Np-3)}{(N-1)(N-2)(N-3)}$$

. . . . . .

§ 6. Selecția dintr-o colectivitate finită

Fie o colectivitate de N elemente, caracterizate fiecare printr-unul dintre numerele

$$a_1,\ a_2,\ldots,a_N$$

diferite între ele.
Notăm

$$M_1=\frac{\sum_{i=1}^{N}a_i}{N},\qquad M_2=\frac{\sum_{i=1}^{N}a_i^2}{N},\ldots,\qquad M_r=\frac{\sum_{i=1}^{N}a_i^r}{N}$$

și

$$m_1=0,\qquad m_2=\frac{\sum_{i=1}^{N}(a_i-m)^2}{N},\ldots,\qquad m_r=\frac{\sum_{i=1}^{N}(a_i-m)^r}{N}\qquad (m=M_1).$$

Scoatem n elemente din această colectivitate de N elemente, fără a pune elementul scos înapoi în colectivitate.
Avem, deci, un șir de variabile vectoriale aleatorii dependente

$$X_1,\ X_2,\ldots,X_i,\ldots,X_n,$$

unde fiecare dintre ele poate lua valorile

$$a_1,\ a_2,\ldots,a_i,\ldots,a_N.$$

Variabila X_i va lua aceste valori cu probabilitățile respective

$$p_1^{(i)},\ p_2^{(i)},\ldots,p_i^{(i)},\ldots,p_N^{(i)}\qquad (i=1,\ 2,\ldots,n),\qquad \left(\sum_{i=1}^{N}p_i=1\right).$$

Să calculăm aceste probabilități. Vom începe cu probabilitatea $p_k^{(i)}$. Presupunem rezultatele extracțiilor succesive dispuse în ordinea

$$\xi_1, \xi_2, \ldots, \xi_i, \ldots, \xi_j, \ldots, \xi_N.$$

Numărul cazurilor posibile este $N!$.
Numărul cazurilor favorabile se obține asociind elementul a_k, apărut în extracția de rangul i cu toate grupele de cîte $N-1$ elemente care au ieșit în celelalte extracții de rang diferit de i.
Acest număr este egal cu $(N-1)!$
Atunci, valoarea probabilității $p_k^{(i)}$ este

$$p_k^{(i)} = \frac{(N-1)!}{N!} = \frac{1}{N},$$

adică independentă de rangul i al extracției.
In cazul acesta, toate probabilitățile $p_k^{(i)} (i=1, 2, \ldots, N)$ sînt egale cu $\frac{1}{N}$.

Fiecare variabilă dă naștere, prin urmare, distribuției

$$X_i \begin{pmatrix} a_1, & a_2, \ldots, a_N \\ \frac{1}{N}, & \frac{1}{N}, \ldots, \frac{1}{N} \end{pmatrix} \qquad (i=1, 2, \ldots, N).$$

Ne propunem să studiem momentele variabilei vectoriale sumă

$$Y_n = X_1 + X_2 + \ldots + X_n.$$

1. *Momentele de ordinul întîi*

$$M_1^{(n)} = M(Y_n) = M\left(\sum_{i=1}^{n} X_i\right) = n\,M(X_i) = n\frac{\sum_{i=1}^{n} a_i}{N} = nM_1. \qquad \text{(IX.6,1)}$$

2. *Momentele de ordinul al doilea*

$$M_2^{(n)} = M(Y_n^2) = M\left(\sum_{i=1}^{n} X_i\right)^2 = \sum_{i=1}^{n} M(X_i^2) + 2\sum_{i \neq j} M(X_i X_j) = n\frac{\sum_{i=1}^{n} a_i^2}{N} +$$

$$+ n(n-1)\,M(X_i X_j). \qquad \text{(IX.6,2)}$$

Valoarea lui $M(X_iX_j)$ o deducem considerînd variabila X_iX_j care va lua, evident, valorile

$$(X_i=a_1,\ X_j=a_1),$$

$$(X_i=a_1,\ X_j=a_2),\ldots,(X_i=a_1,\ X_j=a_N),\ldots,(X_i=a_N,\ X_j=a_N).$$

Pentru determinarea probabilității $P(X_i=a_r,\ X_j=a_s)$, de exemplu, presupunem de asemenea că rezultatele extracțiilor sînt dispuse în ordinea

$$\xi_1,\ \xi_2,\ldots,\xi_i,\ldots,\xi_j,\ldots,\xi_N.$$

Numărul cazurilor posibile este bineînțeles $N!$
Numărul cazurilor favorabile este $2(N-2)!$, deoarece putem avea un element a_r în extracția de rang i și alt element a_s în extracția de rang j, asociind în toate modurile posibile aceste două elemente cu toate grupele de cîte $(N-2)$ elemente apărute în restul extracțiilor.
Această probabilitate va fi

$$P(X_i=a_r,\ X_j=a_s)=\frac{2(N-2)!}{N!}=\frac{2}{N(N-1)},$$

la fel, independentă de i și j.
Expresia (IX.6,2) se poate calcula acum

$$M_2^{(n)}=\frac{n\sum_{i=1}^{n}a_i^2}{N}+\frac{2n(n-1)}{N(N-1)}\sum_{i\neq j}a_ia_j.$$

Se știe că

$$m_2^{(n)}=M_2^{(n)}-[M_1^{(n)}]^2=\frac{n\sum_{i=1}^{n}a_i^2}{N}+\frac{2n(n-1)}{N(N-1)}\sum_{i\neq j}a_ia^j-\frac{n^2\left(\sum_{i=1}^{n}a_i\right)^2}{N^2}.$$

Mai departe vom scrie

$$m_2^{(n)}=n\left[\frac{\sum_{i=1}^{n}a_i^2}{N}+\frac{2(n-1)\sum a_ia_j}{N(N-1)}+\right.$$

$$\left.+\frac{(n-1)\sum_{i=1}^{n}a_i^2}{N(N-1)}-\frac{(n-1)\sum_{i=1}^{n}a_i^2}{N(N-1)}-\frac{n\left(\sum_{i=1}^{n}a_i\right)^2}{N^2}\right]=$$

$$=n\left[\frac{\sum_{i=1}^{n}a_i^2}{N}+\frac{n-1}{N(N-1)}\left(\sum_{i=1}^{n}a_i\right)^2-\frac{n\sum_{i=1}^{n}a_i^2}{N(N-1)}+\frac{\sum_{i=1}^{n}a_i^2}{N(N-1)}-n\left(\frac{\sum_{i=1}^{n}a_i}{N}\right)^2\right]=$$

$$-n\left[\left(\frac{1}{N}-\frac{n}{N(N-1)}+\frac{1}{N(N-1)}\right)\sum_{i=1}^{n}a_i^2-\left(n-\frac{N(n-1)}{N-1}\right)\left(\frac{\sum_{i=1}^{n}a_i}{N}\right)^2\right]=$$

$$=n\left[\frac{N-n}{N-1}\frac{\sum_{i=1}^{n}a_i^2}{N}-\frac{N-n}{N-1}\left(\frac{\sum_{i=1}^{n}a_i}{N}\right)^2\right]=n\frac{N-n}{N-1}\left[\frac{\sum_{i=1}^{n}a_i^2}{N}-\left(\frac{\sum_{i=1}^{n}a_i}{N}\right)^2\right].$$

In sfîrșit, avem

$$m_2^{(n)}=n\frac{N-n}{N-1}m_2.$$

De aici, rezultă următoarea

Teoremă. *Dacă scoatem din colectivitatea totală n elemente la întîmplare, fără să le punem înapoi, dispersia acestei selecții este egală cu dispersia colectivității, înmulțită cu factorul* $\frac{N-n}{N-1}$ *n.*

§ 7. Valoarea asimptotică a probabilității $P_n(\alpha)$

Am stabilit mai înainte valoarea probabilității ca din n extracții succesive efectuate dintr-o urnă, care conține A bile albe și B bile negre ($A+B=N$) să scoatem α bile albe și β bile negre ($\alpha+\beta=n$). După fiecare extracție, bila scoasă nu se mai pune înapoi în urnă. Valoarea acestei probabilități este dată de expresia

$$P_n(\alpha)=\frac{C_A^{\alpha}\cdot C_B^{n-\alpha}}{C_{A+B}^{\alpha+\beta}}=\frac{C_{Np}^{\alpha}\cdot C_{Nq}^{\beta}}{C_N^n}.$$

In aplicații, calculul lui $P_n(\alpha)$ prezintă interes pentru valori mari ale lui n.

Introducem abaterea

$$l=\alpha-np.$$

Inlocuind pe α cu $np+l$, probabilitatea $P_n(\alpha)$ se scrie

$$P_n(l)=\frac{(Np)!}{(np+l)!\,(Np-np+l)!}\cdot\frac{(Nq)!}{(nq-l\,!\,(Nq-nq+l)!}\times$$

$$\times\frac{n!\,(N-n)!}{N!}. \qquad \text{(IX.7,1)}$$

Folosind acum formula asimptotică a lui *Moivre-Stirling*, expresia (IX.7,1) devine

$$P_n(l) \simeq \frac{1}{\sqrt{2\pi}} \cdot \frac{(Np)^{Np+\frac{1}{2}} \cdot (Nq)^{Nq+\frac{1}{2}}}{(np+l)^{np+l+\frac{1}{2}} [(N-n)p-l]^{(N-n)p-l+\frac{1}{2}} (nq-l)^{nq-l+\frac{1}{2}}} \times$$

$$\times \frac{n^{n+\frac{1}{2}} \cdot (N-n)^{N-n+\frac{1}{2}}}{[(N-n)q+l]^{(N-n)q+l+\frac{1}{2}} \cdot N^{N+\frac{1}{2}}} =$$

$$= \frac{\sqrt{N}}{N} \cdot \frac{(np)^{-np-l-\frac{1}{2}} [(N-n)]^{-(N-n)p+l-\frac{1}{2}} (nq)^{-nq+l-\frac{1}{2}}}{\left(1+\frac{l}{np}\right)^{np+l+\frac{1}{2}} \left[1-\frac{l}{(N-n)p}\right]^{(N-n)p-l+\frac{1}{2}}} \times$$

$$\times \frac{[(N-n)q]^{-(N-n)q-l-\frac{1}{2}} \cdot p^{Np+\frac{1}{2}} q^{Nq+\frac{1}{2}} \cdot n^{n+\frac{1}{2}}}{\left(1-\frac{l}{nq}\right)^{nq-l+\frac{1}{2}}} \times$$

$$\times \frac{(N-n)^{N-n+\frac{1}{2}}}{\left[1+\frac{l}{(N-n)q}\right]^{(N-n)q+l+\frac{1}{2}}} = \frac{\sqrt{N}}{\sqrt{2\pi N(N-n)pq}} \cdot \frac{1}{T},$$

$$T = \left(1+\frac{l}{np}\right)^{np+l+\frac{1}{2}} \left[1-\frac{l}{(N-n)p}\right]^{(N-n)p-l+\frac{1}{2}} \times$$

$$\times \left(1-\frac{l}{nq}\right)^{nq-l+\frac{1}{2}} \left[1+\frac{l}{(N-n)q}\right]^{(N-n)q+l+\frac{1}{2}}. \qquad \text{(IX.7,2)}$$

Pentru evaluarea asimptotică a lui T, presupunem că este îndeplinită condiția

$$A\sqrt{n} \leq l \leq B\sqrt{n},$$

A și B fiind două constante.
În acest caz, au loc inegalitățile

$$\frac{|l|}{np}<1,\ \frac{|l|}{(N-n)p}<1,\ \frac{|l|}{nq}<1,\ \frac{|l|}{(N-n)q}<1,$$

pentru valorile mari ale lui n.
Aplicăm expresiei (IX.7,2) logaritmii

$$\ln T=\left(np+l+\frac{1}{2}\right)\ln\left(1+\frac{l}{np}\right)+$$

$$+\left[(N-n)p-l+\frac{1}{2}\right]\ln\left[1-\frac{l}{(N-n)p}\right]+$$

$$+\left(nq-l+\frac{1}{2}\right)\ln\left(1-\frac{l}{nq}\right)+\left[(N-n)q+l+\frac{1}{2}\right]\ln\left[1+\frac{l}{(N-n)q}\right].$$

Dezvoltînd în serie funcțiile logaritmice din membrul al doilea, avem

$$\left(np+l+\frac{1}{2}\right)\ln\left(1+\frac{l}{np}\right)=\left(np+l+\frac{1}{2}\right)\left(\frac{l}{np}-\frac{l^2}{2n^2p^2}+\ldots\right)=$$

$$=l+\frac{l^2}{np}+O\left(\frac{1}{n^2}\right),$$

$$\left[(N-n)p-l+\frac{1}{2}\right]\ln\left[1-\frac{l}{(N-n)p}\right]=$$

$$-\left[(N-n)p-l+\frac{1}{2}\right]\left[-\frac{l}{(N-n)p}-\frac{l^2}{2(N-n)^2p^2}-\ldots\right]=$$

$$=-l+\frac{l^2}{2(N-n)p}+O\left(\frac{1}{n^2}\right),$$

$$\left(nq-l+\frac{1}{2}\right)\ln\left(1-\frac{l}{nq}\right)=\left(nq-l+\frac{1}{2}\right)\left(-\frac{l}{nq}-\frac{l^2}{2nq^2}-\ldots\right)=$$

$$=-l+\frac{l^2}{2nq}+O\left(\frac{1}{n^2}\right).$$

$$\left[(N-n)\,q+l+\frac{1}{2}\right]\ln\left[1+\frac{l}{(N-n)\,q}\right]= \qquad \text{(IX.7,3)}$$

$$=\left[(N-n)\,q+l+\frac{1}{2}\right]\left[\frac{l}{(N-n)\,q}-\frac{l^2}{2\,(N-n)^2q^2}+\ldots\right]=$$

$$=l+\frac{l^2}{2\,(N-n)\,q}+O\left(\frac{1}{n^2}\right).$$

Adunînd egalitățile (IX.7,3), obținem

$$\ln T=\frac{N}{2n\,(N-n)\,pq}\cdot l^2+O\left(\frac{1}{n^2}\right),$$

de unde

$$T=e^{\frac{N}{2n\,(N-n)\,pq}l^2+O\left(\frac{1}{n^2}\right)}.$$

Înlocuind această valoare a lui T în expresia (IX.7,1), găsim

$$P_n(l)\simeq\frac{\sqrt{N}}{\sqrt{2\pi n\,(N-n)\,pq}}\,e^{-\frac{N}{2n\,(N-n)\,pq}l^2}. \qquad \text{(IX.7,4)}$$

Așadar, am demonstrat următoarea:

Teoremă. *Probabilitatea ca să obținem din n extracții succesive, fără a pune bila extrasă înapoi în urnă, o abatere* $A\sqrt{n}\leq l\leq B\sqrt{n}$, *A și B fiind două constante, este asimptotic egală cu*

$$P_n(l)\simeq\frac{\sqrt{N}}{\sqrt{2\pi n\,(N-n)\,pq}}\,e^{-\frac{N}{2n\,(N-n)\,pq}l^2}.$$

Expresia (IX.7,4) a probabilităţii $P_n(l)$ este o lege normală, avînd dispersia σ^2 egală cu

$$\frac{n(N-n)}{N} pq.$$

Valoarea exactă a lui σ^2 a fost calculată mai înainte, cînd ne-am ocupat de momentele centrate ale distribuţiei hipergeometrice. Am găsit atunci pentru această valoare expresia

$$\frac{n(N-n)}{N-1} pq.$$

La limită, pentru $N \to \infty$, ambele valori ale lui σ^2 tind, evident, către dispersia distribuţiei binomiale

$$npq,$$

iar probabilitatea $P_n(l)$ tinde, de asemenea, către aceeaşi lege normală

$$\frac{1}{\sqrt{2\pi\, npq}} e^{-\frac{l^2}{2npq}}$$

către care tinde şi probabilitatea $P_n(l)$ a distribuţiei binomiale.

Procesul de convergenţă din teorema de mai sus este, în general, rapid.

Ne mulţumim să arătăm acest lucru pe exemplul următor (tabela IX.7,1).

De observat că, pe măsură ce numărul n al extracţiilor creşte, tabela probabilităţilor $10\,000\, P_n(l)$ începe să devină simetrică.

Astfel, pentru $n=100$, se vede o simetrie destul de accentuată în jurul valorii medii $np=10$, căreia îi corespunde cea mai mare probabilitate $10\,000\, P_{100}(l) = 1\,319$. De o parte şi de alta a acestei probabilităţi maxime, valorile probabilităţii $10\,000\, P_{100}(l)$ echidistante, deşi diferă între ele, ne dăm seama că se vor apropia din ce în ce mai mult, cu cît numărul n va fi mai mare.

Scanned with OKEN Scanner

TABELA IX.7,1

Valorile probabilității 10 000 $P_n(l)$ în baza schemei bilei neîntoarse. $N=400$; $n=5, 10, 20, 50, 100$; $p=0,1$; $q=0,9$

$\alpha=l+np$	$n=5$	$n=10$	$n=20$	$n=50$	$n=100$
0	5 905	3 487	1 216	51	—
1	3 281	3 874	2 702	286	3
2	729	1 937	2 852	779	16
3	81	574	1 901	1 386	59
4	4	112	898	1 809	159
5		14	319	1 849	338
6		2	89	1 541	596
7			20	1 076	889
8			3	643	1 148
9				333	1 304
10				155	1 319
11				61	1 199
12				22	988
13				7	743
14				2	513
15					327
16					193
17					106
18					54
19					26
20					12
21					5
22					2
23					1

§ 8. Distribuția hipergeometrică multidimensională

Să considerăm următoarea problemă mai generală.

Într-o urnă se găsesc N_1 bile de culoarea întîi, N_2 bile de culoarea a doua, N_3 bile de culoarea a treia, ..., N_m bile de culoarea a m-a ($N_1+N_2+N_3+\ldots+N_m=N$). Din urnă se fac n extracții succesive, fără să se pună bila extrasă înapoi în urnă.

Să se determine probabilitatea $P_n(\alpha_1, \alpha_2, \ldots, \alpha_m)$ ca din n extracții succesive să obținem α_1 bile de culoarea întîi, α_2 bile de culoarea a doua, ... α_m bile de culoarea a m-a ($\alpha_1+\alpha_2+\ldots+\alpha_m=n$).

Această problemă constituie generalizarea firească a schemei tratate în paragraful precedent, cînd aveam în urnă bile numai de două culori: N_1 bile albe şi N_2 bile negre $(N_1+N_2=N)$.
Probabilitatea în acest caz de a scoate din urnă α_1 bile albe şi α_2 bile negre din n extracţii succesive $(\alpha_1+\alpha_2=n)$, fără să se pună bila extrasă înapoi în urnă, a fost egală cu

$$P_n(\alpha_1,\ \alpha_2)=\frac{C_{N_1}^{\alpha_1}\cdot C_{N_2}^{\alpha_2}}{C_{N_1+N_2}^{\alpha_1+\alpha_2}}.$$

Urmînd acelaşi raţionament pe care l-am făcut pentru determinarea probabilităţii $P_n(\alpha_1,\ \alpha_2)$ (cap. IX, § 2), se determină probabilitatea $P_n(\alpha_1,\ \alpha_2,\ldots,\ \alpha_m)$ de a obţine în n extracţii succesive, în baza aceleiaşi scheme probabilistice, α_1 bile de culoarea întîi, α_2 bile de culoarea a doua etc.
Valoarea acestei probabilităţi este

$$P_n(\alpha_1,\ \alpha_2,\ldots,\ \alpha_m)=\frac{C_{N_1}^{\alpha_1}\cdot C_{N_2}^{\alpha_2}\ldots C_{N_m}^{\alpha_m}}{C_{N_1+N_2+\ldots+N_m}^{\alpha_1+\alpha_2+\ldots+\alpha_m}}.$$

Să introducem abaterea redusă,

$$x_h=\frac{\alpha_h-np_h}{\sqrt{\frac{n(N-n)}{N}p_hq_h}}=\frac{\alpha_h-np_h}{\sqrt{\lambda p_hq_h}},\qquad\text{(IX.8,1)}$$

$$\left(\lambda=\frac{n(N-n)}{N}\right),\quad\left(p_h=\frac{N_h}{N}\right),\quad(q_h=1-p_h).$$

Teoremă. *Oricare ar fi* $B_h\leqq x_h\leqq A_h$, *unde* A_h *şi* B_h $(h=1,\ 2,\ldots,\ m)$ *sînt două constante, avem asimptotic*

$$P_n(\alpha_1,\ \alpha_2,\ldots,\ \alpha_m)\simeq\frac{1}{(2\pi)^{\frac{m-1}{2}}\sqrt{p_1p_2\ldots p_m}}\left[\frac{N}{n(N-n)}\right]^{\frac{m-1}{2}}e^{-\frac{1}{2}\sum_{h=1}^{m}q_hx_h^2}.$$

Vom demonstra teorema, servindu-ne de formula asimptotică a lui Moivre-Stirling.
Putem scrie, ţinînd seama de (IX.8,1),

$$P_n(\alpha_1,\ \alpha_2,\ \ldots,\ \alpha_m)=\frac{n!\,(N-n)!}{N!}\prod_{h=1}^{m}\frac{N_h!}{\alpha_h!\,(N_h-\alpha_h)!}\simeq\frac{1}{(2\pi)^{m-1}}\cdot\frac{n^{n+\frac{1}{2}}(N-n)^{N-n+\frac{1}{2}}}{N^{N+\frac{1}{2}}}$$

$$\prod_{h=1}^{m}\frac{(Np_h)^{Np_h+\frac{1}{2}}}{(np_h+\sqrt{\lambda p_h q_h}\cdot x_h)^{np_h+\sqrt{\lambda p_h q_h}\cdot x_h+\frac{1}{2}}\,[(N-n)\,p_h-\sqrt{\lambda p_h q_h}\cdot x_h]^{(N-n)\,p_h-\sqrt{\lambda p_h q_h}\cdot x_h+\frac{1}{2}}}=$$

$$=\frac{N^{\frac{m-1}{2}}\cdot n^{n+\frac{1}{2}}(N-n)^{N-n+\frac{1}{2}}}{(2\pi)^{\frac{m-1}{2}}}\prod_{h=1}^{m}\frac{(np_h)^{-np_h-\sqrt{\lambda p_h q_h}\cdot x_h-\frac{1}{2}}[(N-n)\,p_h]^{-(N-n)\,p_h+\sqrt{\lambda p_h q_h}\cdot x_h-\frac{1}{2}}(p_h)^{Np_h+\frac{1}{2}}}{\left(1+\frac{\sqrt{\lambda p_h q_h}\cdot x_h}{np_h}\right)^{np_h+\sqrt{\lambda p_h q_h}\cdot x_h+\frac{1}{2}}\left[1-\frac{\sqrt{\lambda p_h q_h}\cdot x_h}{(N-n)\,p_h}\right]^{(N-n)\,p_h-\sqrt{\lambda p_h q_h}\cdot x_h+\frac{1}{2}}}=$$

$$=\frac{1}{(2\pi)^{\frac{m-1}{2}}\sqrt{p_1p_2\ldots p_m}}\left[\frac{N}{n\,(N-n)}\right]^{\frac{m-1}{2}}$$

$$\prod_{h=1}^{m}\frac{1}{\left(1+\frac{\sqrt{\lambda p_h q_h}\cdot x_h}{np_h}\right)^{np_h+\sqrt{\lambda p_h q_h}\cdot x_h+\frac{1}{2}}\left[1-\frac{\sqrt{\lambda p_h q_h}\cdot x_h}{(N-n)\,p_h}\right]^{(N-n)\,p_h-\sqrt{\lambda p_h q_h}\cdot x_h+\frac{1}{2}}}=$$

$$=\frac{1}{(2\pi)^{\frac{m-1}{2}}\sqrt{p_1p_2\ldots p_m}}\left[\frac{N}{n\,(N-n)}\right]^{\frac{m-1}{2}}\prod_{h=1}^{m}\frac{1}{T_h}\cdot\qquad\text{(IX.8,2)}$$

Scanned with OKEN Scanner

De remarcat în efectuarea calculelor de mai sus că

$$\sum_{h=1}^{m} \sqrt{\lambda p_h q_h} \cdot x_h = \sum_{h=1}^{m} (\alpha_h - np_h) = \sum_{h=1}^{m} \alpha_h - n \sum_{h=1}^{m} p_h = n - n = 0,$$

$$T_h = \left(1 + \frac{\sqrt{\lambda p_h q_h} \cdot x_h}{np_h}\right)^{np_h + \sqrt{\lambda p_h q_h} \cdot x_h + \frac{1}{2}} \times \qquad \text{(IX.8,3)}$$

$$\times \left[1 - \frac{\sqrt{\lambda p_h q_h} \cdot x_h}{(N-n)\, p_h}\right]^{(N-n)\, p_h - \sqrt{\lambda p_h q_h} \cdot x_h + \frac{1}{2}}.$$

Aplicînd logaritmii în expresia (IX.8,3) obținem

$$\ln T_h = \left(np_h + \sqrt{\lambda p_h q_h} \cdot x_h + \frac{1}{2}\right) \ln \left(1 + \frac{\sqrt{\lambda p_h q_h}}{np_h} x_h\right) +$$

$$+ \left[(N-n)\, p_h - \sqrt{\lambda p_h q_h} \cdot x_h + \frac{1}{2}\right] \ln \left[1 - \frac{\sqrt{\lambda p_h q_h}}{(N-n)\, p_h} x_h\right].$$

Din condițiile teoremei rezultă

$$\left|\frac{\sqrt{\lambda p_h q_h}}{np_h} x_h\right| < 1, \quad \left|\frac{\sqrt{\lambda p_h q_h}}{(N-n)\, p_h} x_h\right| < 1,$$

pentru valori mari ale lui n.

Atunci, dezvoltînd în serie funcțiile logaritmice din expresia lui $\ln T_h$ avem

$$\left(np_h + \sqrt{\lambda p_h q_h} \cdot x_h + \frac{1}{2}\right) \ln \left(1 + \frac{\sqrt{\lambda p_h q_h}}{np_h}\right) =$$

$$= \left(np_h + \sqrt{\lambda p_h q_h} \cdot x_h + \frac{1}{2}\right) \left(\frac{\sqrt{\lambda p_h q_h}}{np_h} x_h - \frac{\lambda p_h q_h}{2n^2 p_h^2} x_h^2 + \ldots\right) =$$

$$= \sqrt{\lambda p_h q_h} \cdot x_h + \frac{\lambda q_h}{2n} x_h^2 + 0(n^{-2}),$$

$$\left[(N-n)\,p_h - \sqrt{\lambda p_h q_h}\cdot x_h + \frac{1}{2}\right] \ln\left[1 - \frac{\sqrt{\lambda p_h q_h}}{(N-n)\,p_h}\,x_h\right] =$$

$$= \left[(N-n)\,p_h - \sqrt{\lambda p_h q_h}\cdot x_h + \frac{1}{2}\right]\left[-\frac{\sqrt{\lambda p_h q_h}}{(N-n)\,p_h}\,x_h - \frac{\lambda p_h q_h}{2(N-n)^2 p_h^2}\,x_h^2 - \dots\right] =$$

$$= -\sqrt{\lambda p_h q_h}\cdot x_h + \frac{\lambda q_h}{2(N-n)}\,x_h^2 + O(n^{-2}). \qquad \text{(IX.8,4)}$$

Adunînd egalitățile (IX. 8,4), găsim

$$\ln T_h = \frac{1}{2}\,q_h x_h^2 + O(n^{-2}),$$

de unde

$$T_h = e^{\frac{1}{2}\,q_h x_h^2 + O(n^{-2})}.$$

Înlocuind în (IX.8,2) această valoare a lui T_h, deducem

$$P_n(\alpha_1, \alpha_2, \dots, \alpha_m) \simeq \frac{1}{(2\pi)^{\frac{m-1}{2}}\sqrt{p_1 p_2 \dots p_m}} \left[\frac{N}{n(N-n)}\right]^{\frac{m-1}{2}} \cdot e^{-\frac{1}{2}\sum_{h=1}^{m} q_h x_h^2}. \qquad \text{(IX.8,5)}$$

Teorema este astfel demonstrată.

Observăm că, pentru $N \to \infty$, expresia asimptotică (IX. 8,5) a probabilității $P_n(\alpha_1, \alpha_2, \dots, \alpha_m)$ tinde către expresia asimptotică a probabilității $P_n(\alpha_1, \alpha_2, \dots, \alpha_m)$ din schema probabilistică a distribuției multinomiale.

In adevăr

$$\lim_{N\to\infty} P_n(\alpha_1, \alpha_2, \dots, \alpha_m) =$$

$$= \lim_{N\to\infty} \frac{1}{(2\pi)^{\frac{m-1}{2}}\sqrt{p_1 p_2 \dots p_m}} \left[\frac{N}{n(N-n)}\right]^{\frac{m-1}{2}} e^{-\frac{1}{2}\sum_{h=1}^{m} q_h x_h^2} =$$

$$= \frac{1}{(2\pi n)^{\frac{m-1}{2}}\sqrt{p_1 p_2 \dots p_m}}\, e^{-\frac{1}{2}\sum_{h=1}^{m} q_h x_h^2}.$$

Explicația rezultă de asemenea, în baza faptului că, avînd în urnă un număr infinit de bile, compoziția urnei nu se schimbă după fiecare din cele n extracții succesive (n finit), dacă nu se pune bila extrasă înapoi în urnă.

§ 9. Momentele distribuției hipergeometrice multidimensionale

Să considerăm din nou o urnă care conține N_1 bile de culoarea întîi, N_2 bile de culoarea a doua, ..., N_m bile de culoarea a m-a. Facem n extracții succesive, fără să punem bila extrasă înapoi în urnă. Fiecare extracție o considerăm ca o variabilă aleatorie care ia valoarea a_1 dacă am extras o bilă de culoarea 1, valoarea a_2 dacă am extras o bilă de culoarea 2, ..., valoarea a_m dacă am extras o bilă de culoarea m.

Prin urmare, avem o succesiune de variabile aleatorii dependente

$$X_1,\ X_2,\ \ldots,\ X_n.$$

Variabila X_i ia valorile

$$a_1,\ a_2,\ \ldots,\ a_m$$

cu probabilitățile respective

$$p_{i_1},\ p_{i_2},\ \ldots,\ p_{i_m} \qquad (i=1,\ 2,\ \ldots,\ n),\quad \left(\sum_{k=1}^{m} p_{i_k}=1\right)$$

care urmează să fie determinate.

Evident avem

$$p_{i_k}=P(X_i=a_k).$$

Vrem să calculăm momentele distribuției, determinată de variabila aleatorie sumă

$$X=X_1+X_2+\ldots+X_n.$$

Notăm

$$X_1+X_2+\ldots+X_n=\lambda_1 a_1+\lambda_2 a_2+\ldots+\lambda_m a_m \qquad \left(\sum_{i=1}^{m}\lambda_i=n\right). \qquad (IX.9,1)$$

1. *Momentele de ordinul întîi*

$$M_1=M(X)=M\left(\sum_{i=1}^{m} X_i\right)=n\sum_{k=1}^{m} p_{i_k}a_k.$$

Ținînd seama de (IX. 9,1), mai putem scrie

$$M_1 = M(X) = M\left(\sum_{k=1}^{m} \lambda_k a_k\right) = \sum_{k=1}^{m} M(\lambda_k)\, a_k.$$

Cele două expresii ale lui M_1 avînd loc, oricare ar fi numerele $a_k\,(k=1,\ 2,\ldots,\ n)$, rezultă imediat

$$M(\lambda_k) = \sum_{i=1}^{n} p_{i_k} \qquad (k=1,\ 2,\ldots,\ m). \qquad \text{(IX. 9,2)}$$

Vom determina probabilitatea $p_{i_k} = P(X_i = a_k)$.

Presupunem că cele N valori posibile, după ce am extras toate bilele din urnă, sînt aranjate în ordinea următoare

$$\xi_1,\ \xi_2,\ldots,\ \xi_i,\ldots,\ \xi_j,\ldots,\ \xi_N.$$

Numărul cazurilor posibile este evident N!

Numărul cazurilor favorabile va fi $N_i\,(N-1)!$, deoarece fiecare valoare a_i din cele N_i se poate scrie cu $(N-1)!$ grupe de cîte $N-1$ valori rămase.

Deci, vom avea

$$p_{i_k} = \frac{N_i\,(N-1)!}{N!} = \frac{N_i}{N} = p_i,$$

adică p_{i_k} este independentă de k.

Atunci, înlocuind pe p_{i_k} cu această valoare în (IX. 9,2), găsim

$$M(\lambda_i) = np_i \qquad (i=1,\ 2,\ldots,\ m).$$

2. *Momentele de ordinul al doilea*

$$M_2 = M(X^2) = M\left(\sum_{i=1}^{n} X_i\right)^2 = \sum_{i=1}^{n} M(X_i^2) +$$

$$+2\sum_{i\neq j} M(X_iX_j) = n\sum_{i=1}^{n} p_i a_i^2 + 2\sum_{i\neq j} M(X_iX_j). \qquad \text{(IX. 9,3)}$$

Pentru determinarea lui M_2, va trebui să calculăm pe $M(X_iX_j)$.

Variabila aleatorie X_iX_j poate lua valorile

$$(X_i=a_1,\ X_j=a_1),\ (X_i=a_1,\ X_j=a_2),\ldots,\ (X_i=a_1,\ X_j=a_m),\ldots$$
$$\ldots,\ (X_i=a_m,\ X_j=a_m)$$

cu probabilităţile respective

$$P(X_i=a_1,\ X_j=a_1),\ P(X_i=a_1,\ X_j=a_2),\ldots,\ P(X_i=a_1,\ X_j=a_n),\ldots$$
$$\ldots,\ P(X_i=a_m,\ X_j=a_m).$$

Vom avea astfel de calculat probabilităţile

$$P(X_i=a_h,\ X_j=a_h),\ P(X_i=a_h,\ X_j=a_k),\qquad (h\neq k=1,\ 2,\ldots,\ m).$$

Vom presupune următoarea succesiune a celor N valori posibile

$$\xi_1,\ \xi_2,\ldots,\ \xi_i,\ldots,\ \xi_j,\ldots,\ \xi_N.$$

Este uşor de văzut, ca şi pînă acum, că numărul cazurilor posibile este $N!$

Să socotim numărul cazurilor favorabile, cînd variabilele X_i şi X_j iau valori egale cu a_h.

Acest număr va fi $N_h(N_h-1)\cdot(N-2)!$, deoarece avem N_h valori egale cu a_h şi ele se pot asocia în toate modurile posibile cu $(N-2)!$ grupuri de cîte $N-2$ valori rămase.

Atunci, vom scrie

$$P(X_i=a_h,\ X_j=a_h)=\frac{N_h(N_h-1)(N-2)!}{N!}=\frac{p_h(Np_h-1)}{N-1}\quad\left(\frac{N_h}{N}=p_h\right).$$

In mod analog, obţinem

$$P(X_i=a_h,\ X_j=a_k)=2\,\frac{N_hN_k(N-2)!}{N!}=2\,\frac{N}{N-1}\cdot p_h\cdot p_k.$$

In acest caz, expresia (IX.9,3) devine

$$M_2=n\sum_{i=1}^{m}p_ia_i^2+\frac{n(n-1)}{N-1}\left[\sum_{h=1}^{m}p_h(N_h-1)\,a_h^2+2\sum_{h\neq k}Np_h\cdot p_k\cdot a_h\cdot a_k\right].$$

Insă, în baza egalităţii (IX.9,1), avem

$$M_2=M(X^2)=M\left(\sum_{i=1}^{m}\lambda_ia_i\right)^2=\sum M(\lambda_i^2)\,a_i^2+2\sum_{i\neq j}M(\lambda_i\lambda_j)\,a_i\cdot a_.$$

De unde deducem, comparînd cele două valori ale lui M_2,

$$\sum_{i=1}^{m}\left[np_i+\frac{n(n-1)}{N-1}p_i(Np_i-1)\right]a_i^2+2\sum_{i\neq j}\frac{n(n-1)N}{N-1}p_ip_ja_ia_j=$$

$$=\sum_{i=1}^{m}M(\lambda_i^2)\,a_i^2+2\sum_{i\neq j}M(\lambda_i\lambda_j)\,a_ia_j. \qquad \text{(IX. 9,4)}$$

Egalitatea (IX. 9,4) fiind satisfăcută pentru orice $a_i\,(i=1,\ 2,\ldots.,\ m)$, se găseşte

$$M(\lambda_i^2)=np_i+\frac{n(n-1)}{N-1}p_i(Np_i-1), \qquad \text{(IX. 9,5)}$$

$$M(\lambda_i\lambda_j)=\frac{n(n-1)N}{N-1}\cdot p_ip_j \qquad (i\neq j=1,\ 2,\ldots,\ m).$$

In acelaşi fel se pot calcula şi momentele de ordin superior lui doi. Acum putem trece la calculul momentelor centrate ale distribuției hipergeometrice cu m stări, deoarece ele rezultă din cunoaşterea momentelor absolute. Avem

$$M(\mu_i)=0,$$

$$M(\mu_i^2)=M(\lambda_i-np_i)^2=np_i+$$

$$+\frac{n(n-1)}{N-1}p_i(Np_i-1)-n^2p_i^2=np_i(1-p_i)\frac{N-n}{N-1}, \qquad \text{(IX. 9,6)}$$

$$M(\mu_i\mu_j)=M[(\lambda_i-np_i)(\lambda_j-np_j)]=$$

$$=\frac{n(n-1)N}{N-1}p_ip_j-n^2p_ip_j=-\frac{n(N-n)}{N-1}p_ip_j.$$

Tot astfel se vor calcula şi momentele centrate de ordin mai mare. Făcînd pe N să tindă către infinit în expresiile (IX. 9,5) şi (IX. 9,6), regăsim, după cum este şi natural, momentele obişnuite şi momentele centrate de ordinul al doilea ale distribuției mul inomiale.

In adevăr,

$$\lim_{N\to\infty}M(\lambda_i^2)=\lim_{N\to\infty}\left[np_i+\frac{n(n-1)}{N-1}p_i(Np_i-1)\right]=n^2p_i^2+np_i(1-p_i),$$

$$\lim_{N\to\infty}M(\lambda_i\lambda_j)=\lim_{N\to\infty}\left[\frac{n(n-1)N}{N-1}p_ip_j\right]=n(n-1)p_ip_j,$$

$$\lim_{N\to\infty} M(\mu_i^2) = \lim_{N\to\infty}\left[np_i(1-p_i)\frac{N-n}{N-1}\right] = np_i(1-p_i),$$

$$\lim_{N\to\infty} M(\mu_i\mu_j) = \lim_{N\to\infty}\left[-\frac{n(N-n)}{N-1}p_ip_j\right] = -np_ip_j.$$

Capitolul X

Distribuții Poisson. Distribuția binomială cu exponent negativ

§ 1. Considerații generale

Știm că probabilitatea ca un eveniment E cu probabilitatea constantă p în n probe succesive independente să se verifice de α ori este dată de expresia

$$P_n(\alpha) = C_n^\alpha p^\alpha q^{n-\alpha}. \qquad (X.1,1)$$

Pentru valori mari ale lui n, care este de altfel cazul cel ma frecvent și interesant în aplicațiile statisticii, probabilitatea $P_n(\alpha)$ a fost pusă, folosind formula asimptotică a lui Moivre-Stirling, sub o formă apropiată legii normale a lui Laplace.

Pentru realizarea acestui rezultat este esențială condiția ca p să fie constant, nu numai de la probă la probă pentru un sistem de n probe, dar și atunci cînd n variază.

Dacă însă p variază cînd n variază în înțelesul următor: pentru orice sistem de n probe $p=p(n)$ este același, dar dacă considerăm un sistem de $(n+k)$ probe, probabilitatea este diferită, $p=p(n+m)$. Așa, de exemplu, dacă p este foarte mic și n suficient de mare, astfel ca $np=\lambda$ (o constantă), atunci expresia (X.1,1) ne dă, pentru $n\to\infty$, o lege total diferită de cea normală.

In adevăr

$$P_n(\alpha) = \frac{n(n-1)\ldots(n-\alpha+1)}{\alpha!}p^\alpha\cdot q^{n-\alpha} =$$

$$= \frac{n(n-1)\ldots(n-\alpha+1)}{\alpha!}\cdot\frac{\lambda^\alpha}{n^\alpha}\left(1-\frac{\lambda}{n}\right)^{n-\alpha}.$$

Cum însă

$$\lim_{n\to\infty}\frac{n(n-1)\dots(n-\alpha+1)}{n^\alpha}=1 \quad \text{şi} \quad \lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{n-\alpha}=e^{-\lambda},$$

putem scrie

$$P_n(\alpha)\simeq\frac{\lambda^\alpha}{\alpha!}e^{-\lambda} \qquad (\alpha=0,\ 1,\dots,\ \infty). \qquad \text{(X.1,2)}$$

Observăm că expresia $\frac{\lambda^\alpha}{\alpha!}e^{-\lambda}$ este o funcţie de frecvenţă, deoarece avem evident

$$\sum_{\alpha=0}^{\infty}\frac{\lambda^\alpha}{\alpha!}e^{-\lambda}=e^{-\lambda}\sum_{\alpha=0}^{\infty}\frac{\lambda^\alpha}{\alpha!}=1.$$

Vom spune că o distribuţie, care are ca funcţie de frecvenţă expresia $\frac{\lambda^\alpha}{\alpha!}e^{-\lambda}$ este o *distribuţie Poisson*, după numele lui Poisson, care a găsit-o pentru prima oară.

Expresia asimptotică (X.1,2) a probabilităţii $P_n(\alpha)$ ne dă posibilitatea de a întocmi diferite statistici, în care nu se aplică repartiţia normală.

Distribuţia Poisson se întîlneşte, după cum am văzut, în cazul evenimentelor care se întîmplă rar, din care cauză această distribuţie mai este cunoscută sub numele de *legea evenimentelor rare*.

L. Bortkievici a folosit de aproape distribuţia Poisson şi a dat numeroase exemple de asemenea distribuţii.

Vom arăta aici cîteva din exemplele lui Bortkievici.

Exemplul 1. Intre anii 1886—1894 s-a făcut o statistică a accidentelor mortale în unsprezece întreprinderi din Germania, în care numărul accidentelor mortale era foarte redus.

Această statistică este dată de tabela X.1,1.

Distribuţia acestor accidente mortale după frecvenţa lor reală şi teoretică este dată în tabela X.1,2.

Valoarea medie a numărului de accidente mortale este egală cu $\lambda=4{,}55$.

In coloana a treia a tabelei (X.1,2) s-au calculat funcţiile de frecvenţă pentru distribuţia Poisson cu parametrul $\lambda=4{,}55$.

După cum se constată, există o concordanţă destul de bună între frecvenţe şi probabilităţi, calculate cu ajutorul expresiei (X.1,2).

Această concordanţă este şi mai strînsă, dacă grupăm frecvenţele după cum se vede în tabela X.1,3.

TABELA X.1,1

Numărul de accidente mortale între anii 1886—1894 în unsprezece întreprinderi

Intreprinderea nr.	1886	1887	1888	1889	1890	1891	1892	1893	1894
13	6	8	7	5	14	8	9	4	8
14	2	2	2	1	1	3	5	3	4
12	—	1	2	2	5	—	2	7	4
20	3	3	5	3	10	2	5	4	4
23	3	9	6	11	6	8	5	4	4
27	1	2	2	3	1	1	1	4	2
29	4	8	4	3	8	3	7	4	12
40	2	5	1	3	2	—	—	6	7
41	1	3	5	4	6	7	5	8	7
42	5	6	5	5	3	4	—	8	5
55	5	5	2	2	7	5	3	6	4

TABELA X.1,2

Distribuţia numărului de cazuri observate în raport cu numărul accidentelor mortale

Numărul accidentelor mortale (1)	Frecvenţe (numărul cazurilor observate) (2)	$P_n(a)$ (3)	Numărul accidentelor mortale (1)	Frecvenţe (numărul cazurilor observate) (2)	$P_n(a)$ (3)
0	5	3,69	9	2	3,30
1	9	9,61	10	1	2,03
2	14	13,89	11	1	1,19
3	13	15,21	12	1	0,66
4	14	14,34	13	—	0,34
5	16	12,31	14	1	0,17
6	7	9,80	15	—	0,08
7	7	7,28	16	—	0,04
8	8	5,06			

Exemplul 2. Intre anii 1875—1894 s-a înregistrat numărul soldaţilor morţi din cauza loviturilor de copite ale cailor în Germania.

Distribuţia acestor decese după frecvenţa lor reală şi teoretică este arătată în tabela X.1,4.

Şi în acest exemplu se constată o concordanţă satisfăcătoare între frecvenţe şi probabilităţi, care s-au calculat de asemenea cu ajutorul expresiei asimptotice (X.1,2) a probabilităţii $P_n(a)$.

TABELA X.1,3

Distribuţia numărului de cazuri observate, în raport cu numărul accidentelor mortale, grupate pe intervale neegale

Numărul accidentelor mortale (1)	Frecvenţe (numărul cazurilor observate) (2)	$P_n(a)$ (3)
0—2	28	27,2
3—4	27	29,6
5—6	23	22,1
7—	21	20,1

TABELA X.1,4

Distribuţia numărului de cazuri observate în raport cu numărul morţilor

Numărul deceselor (1)	Frecvenţe (numărul cazurilor observate) (2)	$P_n(a)$ (3)
0	109	108,7
1	65	66,3
2	22	20,2
3	3	4,1
4	1	0,6
5	0	0,1

Distribuţiile Poisson sînt întîlnite în mod frecvent la controlul statistic al produselor industriale, unde probabilitatea de a se produce un fabricat necorăspunzător este foarte mică.

Distribuţia Poisson se aplică în deosebit de favorabile condiţii în telefonie.

§ 2. Momente

Considerînd funcţia generatoare a momentelor factoriale în cazul distribuţiei Poisson, avem

$$\Phi(t) = M(t^a) = \sum_{a=0}^{\infty} t^a \frac{\lambda^a}{a!} e^{-\lambda} = e^{\lambda(t-1)}.$$

Derivînd funcţia $\Phi(t)$ succesiv de h ori în raport cu t şi făcînd apoi pe $t=1$, obţinem

$$M_{(h)} = \frac{d^h}{dt^h}\left[e^{\lambda(t-1)}\right]_{t=1} = \lambda^h \qquad (h=1,\ 2,\ \ldots).$$

Celelalte momente, obişnuite şi centrate, se calculează imediat cu ajutorul momentelor factoriale.

De exemplu, momentele M_h $(h=1,2,\ldots)$ se obţin, cunoscînd că ele sînt funcţii liniare în raport cu numerele lui Stirling de speţa a doua (Cap. II, § 6)

$$M_h = T_h^1 M_{(1)} + T_h^2 M_{(2)} + \ldots + T_h^h M_{(h)}.$$

Rezultă, deci, următoarele valori pentru primele patru momente obişnuite

$$M_1 = M_{(1)} = \lambda,$$

$$M_2 = M_{(1)} + M_{(2)} = \lambda + \lambda^2,$$

$$M_3 = M_{(1)} + 3M_{(2)} + M_{(3)} = \lambda + 3\lambda^2 + \lambda^3,$$

$$M_4 = M_{(1)} + 7M_{(2)} + 6M_{(3)} + M_{(4)} = \lambda + 7\lambda^2 + 6\lambda^3 + \lambda^4,$$

. . . . . .

Acum, momentele centrate m_h $(h = 1, 2, \ldots)$ se deduc uşor

$$m_1 = 0,$$

$$m_2 = M_2 - M_1^2 = \lambda.$$

$$m_3 = M_3 - 3M_2M_1 + 2M_1^3 = \lambda,$$

$$m_4 = M_4 - 4M_3M_1 + 6M_2M_1^2 - 3M_1^4 = \lambda + 3\lambda^2,$$

. . . . . .

Este interesant de observat că momentele M_1, m_2 şi m_3 ale distribuţiei Poisson sînt egale, valoarea lor comună fiind λ.

§ 3. Funcţia caracteristică

Funcţia caracteristică a distribuţiei Poisson rezultă imediat din definiţia ei

$$\varphi(t, \lambda) = M(e^{\alpha ti}) = \sum_{\alpha=0}^{\infty} e^{\alpha ti} \frac{\lambda^\alpha}{\alpha!} e^{-\lambda} = e^{\lambda(e^{ti}-1)}.$$

Din această expresie a funcţiei caracteristice, rezultă unele consecinţe importante, care se exprimă în teoremele următoare.

Teoremă. *Dacă X şi Y sînt două variabile aleatorii independente, fiecare dintre ele dînd naştere la distribuţia Poisson cu parametrii respectivi λ şi μ, atunci suma lor $X+Y$ urmează, de asemenea, o lege Poisson cu parametrul $\lambda+\mu$.*

Avem

$$P(X=\alpha) = \frac{\lambda^\alpha}{\alpha!} e^{-\lambda}; \quad P(Y=\alpha) = \frac{\mu^\alpha}{\alpha!} e^{-\mu} \quad (\alpha = 0, 1, \ldots, \infty).$$

Atunci, funcţia caracteristică a variabilei $X+Y$, în baza proprietăţii funcţiei caracteristice a sumelor de variabile independente, va fi

$$\varphi(t,\lambda)\cdot\varphi(t,\mu)=e^{\lambda(e^{ti}-1)}\cdot e^{\mu(e^{ti}-1)}=e^{(\lambda+\mu)(e^{ti}-1)}=\varphi(t,\ \lambda+\mu).$$

Deci teorema este demonstrată.

Această teoremă pune în evidenţă o proprietate remarcabilă a distribuţiei Poisson, şi anume, că este o *distribuţie stabilă.*

Teoremă. *Dacă parametrul λ al unei distribuţii Poisson creşte necontenit, atunci legea Poisson a variabilei normată convenabil tinde către legea normală a lui Laplace.*

Se introducem abaterea redusă

$$\frac{\alpha-\lambda}{\sqrt{\lambda}}\qquad(\alpha=0,1,\ldots,\infty)$$

şi să considerăm funcţia caracteristică a acestei abateri.

Vom scrie

$$\psi(t,\lambda)=M\left(e^{\frac{\alpha-\lambda}{\sqrt{\lambda}}ti}\right)=e^{-\sqrt{\lambda}ti}M\left(e^{\alpha\frac{t}{\sqrt{\lambda}}i}\right)=e^{-\sqrt{\lambda}ti}\cdot\varphi\left(\frac{t}{\sqrt{\lambda}},\lambda\right)=$$

$$=e^{-\sqrt{\lambda}ti}\cdot e^{\lambda\left(e^{\frac{t}{\sqrt{\lambda}}i}-1\right)}=e^{\lambda\left(e^{\frac{t}{\sqrt{\lambda}}i}-1\right)-\sqrt{\lambda}ti}=$$

$$=\exp\left\{\lambda\left(e^{\frac{t}{\sqrt{\lambda}}i}-1\right)-\sqrt{\lambda}ti\right\}=\exp\left\{\lambda\left[\frac{t}{\sqrt{\lambda}}i-\frac{t^2}{2\lambda}+\right.\right.$$

$$\left.\left.+O\left(\lambda^{-\frac{3}{2}}\right)\right]-\sqrt{\lambda}ti\right\}=\exp\left\{-\frac{t^2}{2}+O\left(\lambda^{-\frac{1}{2}}\right)\right\},$$

de unde, deducem

$$\lim_{\lambda\to\infty}\psi(t,\lambda)=e^{-\frac{t^2}{2}},$$

care este funcţia caracteristică a legii normale cu dispersia $\sigma^2=1$.

Dacă funcţiile caracteristice $\psi(t,\lambda)$ tind către funcţia caracteristică a legii normale, atunci se ştie că funcţiile de repartiţie corespunzătoare vor tinde de asemenea către funcţia de repartiţie a legii inormale.

Teorema este deci demonstrată.

Scanned with OKEN Scanner

§ 4. Distribuţia Poisson multidimensională

Se dă şirul de variabile vectoriale aleatorii independente

$$X_1, X_2, \ldots, X_n,$$

unde fiecare dintre aceste variabile poate lua valorile

$$a_1, a_2, \ldots, a_m$$

cu probabilităţile respective

$$p_1, p_2, \ldots, p_m \quad \left(\sum_{k=1}^{m} p_k = 1\right),$$

unde presupunem că $p_1, p_2, \ldots, p_m$ sînt foarte mici şi în consecinţă $p_m = 1 - \sum_{k=1}^{m-1} p_k$ este foarte apropiat de 1.

Mai precis, fiecare variabilă vectorială $X_i\,(i=1,2,\ldots,n)$ dă naştere unei distribuţii

$$X_i \begin{pmatrix} e_1(1,0,\ldots,0), & e_2(0,1,0,\ldots,0), \ldots, & e_m(0,\ldots,0,1) \\ p_1 & , \quad p_2 & , \ldots, \quad p_m \end{pmatrix},$$

vectorii $e_k(0,\ldots,0,\ k,0,\ldots,0)$ fiind liniar independenţi.

Să notăm

$$p_1 = \frac{\lambda_1}{n}, \quad p_2 = \frac{\lambda_2}{n}, \ldots, p_{m-1} = \frac{\lambda_{m-1}}{n},$$

unde $\lambda_1, \lambda_2, \ldots, \lambda_{m-1}$ sînt nişte constante.

Teoremă. *Probabilitatea ca în n probe succesive vectorii* $e_1(1,0,\ldots,0)$, $e_2(0,1,0,\ldots,0), \ldots, e_m(0,0,\ldots,1)$ *să apară respectiv de* α_1 *ori,* α_2 *ori,* $\ldots, \alpha_m$ *ori se comportă asimptotic cum urmează*

$$P_n(\alpha_1, \alpha_2, \ldots, \alpha_m) \simeq \frac{\lambda_1^{\alpha_1} \cdot \lambda_2^{\alpha_2} \cdots \lambda_{m-1}^{\alpha_{m-1}}}{\alpha_1!\,\alpha_2! \ldots \alpha_{m-1}!}\, e^{-\sum_{k=1}^{m-1} \lambda_k} \qquad \text{(X.4,1)}$$

$$\left(\sum_{k=1}^{m} \alpha_k = n\right).$$

Am arătat (cap. VIII, §2) că

$$P_n(\alpha_1, \alpha_2, \ldots, \alpha_m) = \frac{n!}{\alpha_1!\,\alpha_2! \ldots \alpha_m!} p_1^{\alpha_1} p_2^{\alpha_2} \ldots p_m^{\alpha_m}.$$

Ținînd seama de notațiile noastre, avem

$$P_n(\alpha_1, \alpha_2, \ldots, \alpha_m) = \frac{n!}{\alpha_1!\,\alpha_2! \ldots \alpha_{m-1}!} \cdot \frac{\lambda_1^{\alpha_1} \cdot \lambda_2^{\alpha_2} \ldots \lambda_{m-1}^{\alpha_{m-1}}}{\alpha_m!\, n^{n-\alpha_m}} \left(1 - \frac{\sum_{k=1}^{m-1} \lambda_k}{n}\right)^{\alpha_m}.$$

Trecînd acum la limită

$$\lim_{n\to\infty} P_n(\alpha_1, \alpha_2, \ldots, \alpha_m) =$$

$$= \lim_{n\to\infty} \left[\frac{n!}{\alpha_1!, \alpha_2! \ldots \alpha_{m-1}!} \cdot \frac{\lambda_1^{\alpha_1} \cdot \lambda_2^{\alpha_2} \ldots \lambda_{m-1}^{\alpha_{m-1}}}{\alpha_m!\, n^{n-\alpha_m}} \left(1 - \frac{\sum_{k=1}^{m-1} \lambda_k}{n}\right)^{\alpha_m} \right] =$$

$$= \frac{\lambda_1^{\alpha_1} \cdot \lambda_2^{\alpha_2} \ldots \lambda_{m-1}^{\alpha_{m-1}}}{\alpha_1!\,\alpha_2! \ldots \alpha_{m-1}!} e^{-\sum_{k=1}^{m-1} \lambda_k},$$

deoarece

$$\lim_{n\to\infty} \frac{n!}{\alpha_m!\, n^{n-\alpha_m}} = 1, \quad \lim_{n\to\infty} \left(1 - \frac{\sum_{k=1}^{m-1} \lambda_k}{n}\right)^{\alpha_m} = e^{-\sum_{k=1}^{m-1} \lambda_k}.$$

Teorema este, prin urmare, demonstrată.

Observăm că expresia asimptotică (X.4,1) a probabilității $P_n(\alpha_1, \alpha_2, \ldots, \alpha_m)$ se poate pune sub forma unui produs de $m-1$ legi de probabilitate Poisson cu parametrii respectivi $\lambda_1, \lambda_2, \ldots, \lambda_{m-1}$, adică

$$P_n(\alpha_1, \alpha_2, \ldots, \alpha_m) \simeq \prod_{k=1}^{m-1} \frac{\lambda_k^{\alpha_k}}{\alpha_k!} e^{-\lambda_k}.$$

De unde rezultă proprietatea foarte importantă că $\alpha_1, \alpha_2, \ldots, \alpha_{m-1}$ sînt independente între ele. Folosind această proprietate, calculul momentelor distribuției Poisson multidimensionale se va ușura foarte mult.

§ 5. Funcţia caracteristică a distribuţiei Poisson multidimensionale

Considerînd distribuţia

$$X_i\begin{pmatrix} e_1(1,0,\dots,0), e_2(0,1,0,\dots,0), \dots, e_m(0,\dots,0,1) \\ p_1 \quad , \quad p_2 \quad , \dots , \quad p_m \end{pmatrix} (i=1,2,\dots,n), \quad \left(\sum_{k=1}^{m} p_k = 1\right)$$

unde probabilităţile p_k $(k=1,2,\dots,n-1$, sînt foarte mici şi deci $p_m = 1 - \sum_{k=1}^{m-1} p_k$ este foarte apropiat de unu, probabilitatea ca în n probe vectorii $e_1(1,0,\dots,0)$, $e_2(0,1,0,\dots,0)$, $\dots$, $e_m(0,0,\dots,0,1)$ să apară respectiv de α_1 ori, α_2 ori, $\dots$ α_m ori este

$$P_n(\alpha_1,\alpha_2,\dots,\alpha_m) =$$

$$= \frac{n!}{\alpha_1!\,\alpha_2!\dots\alpha_m!}\left(\frac{\lambda_1}{n}\right)^{\alpha_1}\cdot\left(\frac{\lambda_2}{n}\right)^{\alpha_2}\dots\left(\frac{\lambda_{m-1}}{n}\right)^{\alpha_{m-1}}\cdot\left(1-\frac{\sum_{k=1}^{m-1}\lambda_k}{n}\right)^{\alpha_m},$$

$\lambda_1,\lambda_2,\dots,\lambda_m$ fiind nişte constante.

Prin definiţie, funcţia caracteristică corespunzătoare variabilei ca în cele n probe, vectorii $e_1(1,0,\dots,0)$, $e_2(0,1,0,\dots,0)$, $\dots$, $e_m(0,\dots,0,1)$ să apară respectiv de α_1 ori, α_2 ori, $\dots$, α_{m-1} ori se va scrie

$$\varphi(u_1,u_2,\dots,u_{m-1},\lambda_1,\lambda_2,\dots,\lambda_{m-1}) = \sum_{\substack{0 \\ \alpha_1,\alpha_2,\dots,\alpha_m}}^{n} \frac{n!}{\alpha_1!\alpha_2!\dots\alpha_m!}\left(\frac{\lambda_1}{n}\right)^{\alpha_1}\left(\frac{\lambda_2}{n}\right)^{\alpha_2}\dots$$

$$\dots\left(\frac{\lambda_{m-1}}{n}\right)^{\alpha_{m-1}}\left(1-\frac{\sum_{k=1}^{m-1}\lambda_k}{n}\right)^{\alpha_m}\cdot e^{(\alpha_1u_1+\alpha_2u_2+\dots+\alpha_{m-1}u_{m-1})i} =$$

$$= \sum_{\substack{0 \\ \alpha_1,\alpha_2,\dots,\alpha_m}}^{n} \frac{n!}{\alpha_1!\,\alpha_2!\dots\alpha_m!}\left(\frac{\lambda_1}{n}e^{u_1i}\right)^{\alpha_1}\left(\frac{\lambda_2}{n}e^{u_2i}\right)^{\alpha_2}\dots\left(\frac{\lambda_{m-1}}{n}e^{u_{m-1}i}\right)^{\alpha_{m-1}}\times$$

$$\times\left(1-\frac{\sum_{k=1}^{m-1}\lambda_k}{n}\right)^{\alpha_m} = \left[\frac{\lambda_1}{n}e^{u_1i}+\frac{\lambda_2}{n}e^{u_2i}+\dots+\frac{\lambda_{m-1}}{n}e^{u_{m-1}i}+\right.$$

$$\left.+\left(1-\frac{\sum_{k=1}^{m-1}\lambda_k}{n}\right)^{\alpha_m}\right]^n = \left[1+\frac{\sum_{k=1}^{m-1}\lambda_k(e^{u_ki}-1)}{n}\right]^n.$$

Scanned with OKEN Scanner

De unde, rezultă

$$\varphi(u_1, u_2, \ldots, u_{m-1}; \lambda_1, \lambda_2, \ldots, \lambda_{m-1}) \simeq e^{\sum_{k=1}^{m-1} \lambda_k (e^{u_k i}-1)} = \prod_{k=1}^{m-1} e^{\lambda_k (e^{u_k i}-1)},$$

pentru valori suficient de mari ale lui n.
Aceasta este funcția caracteristică a legii Poisson cu $m-1$ dimensiuni.

Din expresia funcției caracteristice a legii Poisson cu $m-1$ dimensiuni, rezultă, de asemenea, unele consecințe care sînt arătate în teoremele de mai jos.

Teoremă. *Dacă X și Y sînt două variabile vectoriale aleatorii independente, fiecare dintre ele urmînd o lege Poisson cu $m-1$ dimensiuni, suma lor $X+Y$ urmează de asemenea o lege Poisson cu $m-1$ dimensiuni.*

Teorema rezultă imediat, scriind că funcția caracteristică a variabilei $X+Y$ este egală cu produsul funcțiilor caracteristice ale variabilelor X și Y

$$\varphi(u_1, u_2, \ldots, u_{m-1}; \lambda_1, \lambda_2, \ldots, \lambda_{m-1}) \cdot \varphi(u_1, u_2, \ldots, u_{m-1}; \mu_1, \mu_2, \ldots, \mu_{m-1}) = e^{\sum_{k=1}^{m-1} (\lambda_k+\mu_k)(e^{u_k i}-1)} = \varphi(u_1, u_2, \ldots, u_{m-1}; \lambda_1+\mu_1, \lambda_2+\mu_2, \ldots, \lambda_{m-1}+\mu_{m-1}).$$

După cum se vede, se păstrează ca și în cazul legii Poisson unidimensionale, aceeași proprietate deosebită, care arată că distribuția Poisson multidimensională este o *distribuție stabilă*.

Teoremă. *Dacă fiecare dintre parametrii λ_i $(i=1, 2, \ldots, m-1)$ ai unei legi Poisson cu $m-1$ dimensiuni cresc necontenit, atunci această lege convenabil normată tinde către legea normală cu $m-1$ dimensiuni.*
Introducem vectorul

$$\lambda\left(\frac{\alpha_1-\lambda_1}{\sqrt{\lambda_1}}, \frac{\alpha_2-\lambda_2}{\sqrt{\lambda_2}}, \ldots, \frac{\alpha_{m-1}-\lambda_{m-1}}{\sqrt{\lambda_{m-1}}}\right)$$

și scriem funcția caracteristică corespunzătoare.

Avem, ţinînd seama de expresia funcţiei caracteristice a legii Poisson cu o singură dimensiune

$$\psi(u_1\, u_2,, \ldots, u_{m-1};\quad \lambda_1, \lambda_2, \ldots, \lambda_{m-1}) = \prod_{k=1}^{m-1} e^{-\sqrt{\lambda_k} u_k i} \left[e^{\lambda_k (e^{u_k i}-1)}\right].$$

Trecînd la limită, obţinem

$$\lim_{\lambda_1, \lambda_2, \ldots, \lambda_{m-1} \to \infty} \psi(u_1, u_2, \ldots, u_{m-1};\ \lambda_1, \lambda_2, \ldots, \lambda_{m-1}) =$$

$$= \lim_{\lambda_1, \lambda_2, \ldots, \lambda_{m-1} \to \infty} \prod_{k=1}^{m-1} e^{-\sqrt{\lambda_k}} \left[e^{\lambda_k (e^{u_k i}-1)}\right] = e^{-\frac{1}{2}\sum_{k=1}^{m-1} u_k^2},$$

adică o lege normală cu $m-1$ dimensiuni.
De observat că

$$\lim_{\lambda_k \to \infty} e^{-\sqrt{\lambda_k}} \left[e^{\lambda_k (e^{u_k i}-1)}\right] = e^{-\frac{1}{2} u_k^2}.$$

Teorema este demonstrată, deoarece funcţiile caracteristice $\psi(u_1, \ldots$ $\ldots, u_{m-1};\ \lambda_1, \ldots, \lambda_{m-1})$ tinzînd către funcţia caracteristică a legii normale cu $m-1$ dimensiuni, rezultă că şi funcţiile de repartiţie corespunzătoare tind, de asemenea, către funcţia de repartiţie a legii normale cu $m-1$ dimensiuni.

§ 6. O altă generalizare a distribuţiei Poisson

Să considerăm un lanţ Markov simplu, constant şi finit. Insemnăm cu $x_i\,(i=1,2,\ldots,)$ variabilele lanţului, $a_i\,(i=1,2,\ldots,m)$ valorile luate de aceste variabile şi p_{ik} probabilitatea de trecere de la a_i la a_k.

$$\sum_k p_{ik} = 1, \quad p_{ik} \geqq 0.$$

Pentru un asemenea lanţ, suma

$$y_n = \frac{\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} M(x_i)}{\sqrt{M\left|\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} M(x_i)\right|^2}}, \qquad (X.6,1)$$

unde $M(Y)$ reprezintă valoarea medie a lui Y, urmează la limită legea normală a lui Laplace, în afară de unele cazuri care fac excepție.

Această convergență pentru unele lanțuri este foarte lentă, în așa măsură încît, pentru valori foarte mari ale lui n, distribuția lui y diferă esențial de legea ei limită. In general, pentru fenomenele din categoria de mai sus, valorile lui n, pentru care există diferența arătată, corespund ordinului de mărime pentru care în practică se pot întocmi în mod efectiv statistici. Legea normală a lui Laplace reprezintă de aceea o limită inaccesibilă de experiențe pentru valorile lui n și este necesar să se întreprindă un studiu pentru stabilirea legilor de probabilitate corespunzătoare distribuțiilor corespunzătoare acestor valori ale lui n suficient de mari.

Am văzut că, pentru evenimentele independente, lenta convergență a frecvenței către legea lui Laplace se datorește valorii mici a probabilității de realizare a evenimentului, care se repetă în probele consecutive.

Am obținut astfel distribuția lui Poisson.

Vom arăta în cele ce urmează că același fenomen se petrece și pentru lanțurile lui Markov, pentru valori foarte mari ale lui n în cazul în care anumite probabilități de trecere sînt foarte mici și deci cînd realizarea evenimentului este rară.

Insemnînd prin $F_k(n,t)$ funcția caracteristică a variabilei

$$x_{(n)} = \sum_{i=1}^{n} x_i,$$

în ipoteza că

$$x_1 = a_h,$$

această funcție este dată de relația

$$F_h(n,t) = \sum_{h=1}^{m} \sum_{j=1}^{s} \frac{1}{(m_j-1)!} D_\lambda^{m_j-1} \frac{\lambda^n G_{hk}(\lambda,t)}{\psi_j(\lambda)}. \qquad \text{(X.6,2)}$$

1) $G_{hk}(\lambda,t)$ reprezintă complementul algebric al elementului cuprins în linia de ordinul k și coloana de ordinul h a determinantului

$$D(\lambda,t) = \| p_{ik} e^{\alpha_{hk} t i} - \lambda \varepsilon_{hk} \|$$
$$\alpha_{hk} = 0, \quad \varepsilon_{hk} = 1 \qquad (i = k), \qquad \text{(X.6,3)}$$
$$\alpha_{hk} = 1, \quad \varepsilon_{hk} = 0 \qquad (i \neq k).$$

Scanned with OKEN Scanner

2) $\lambda_1, \lambda_2, \ldots, \lambda_s$ sînt rădăcinile ecuaţiei

$$D(\lambda, t)=0$$

cu ordinele respective de multiplicitate $m_1, m_2, \ldots, m_s$.

3) $$\psi_j=\frac{D(\lambda, t)}{(\lambda-\lambda_j)^{m_j}}.$$

4) $D_\lambda^m f(\lambda)$ reprezintă derivata de ordinul m a funcţiei $f(\lambda)$ în raport cu λ.

Avem întotdeauna

$$|\lambda_j(t)| \leqslant 1 \qquad (t \neq 0),$$

egalitatea avînd loc numai cînd sînt îndeplinite condiţiile necesare şi suficiente pentru ca valorea medie pătratică a variabilei $x_{(n)}$ să fie asimptotică de un ordin mai mic decît n.

Există cel puţin o rădăcină $\lambda_1(t)$ pentru care

$$\lambda_1(0)=1.$$

Distribuţia normală are loc cînd

$$\lambda_1(0)=1$$

este rădăcină simplă a ecuaţiei

$$D(\lambda, 0)=0.$$

Să examinăm cazul cînd, afară de rădăcina simplă $\lambda_1(0)$, ecuaţia

$$D(\lambda, 0)=0$$

nu admite alte rădăcini egale în modul cu 1.

Funcţia $F_h(n, t)$ devine

$$F_h(n, t)=\frac{\sum_{k=1}^{m} G_{hk}(\lambda_1, t)}{\psi_j(\lambda_1)}\lambda_1^n+\Phi(t, n), \qquad \text{(X.6,4)}$$

$\lambda_1=\lambda_1(t)$ devenind egal cu unu pentru $t=0$, iar $\Phi(t, n)$ tinzînd către zero pentru $n \to \infty$.

Pentru determinarea legii de probabilitate a numărului de apariţii ale variabilei a_1, vom face în formulele de mai sus

$$a_1=1, \quad a_i=0 \qquad (i \neq 1).$$

Presupunem că probabilitățile de trecere p_{hk} sînt funcții de un parametru τ, care ia valori mici, adică avem

$$p_{hk}=p_{hk}(\tau).$$

In acest caz, funcția

$$\lambda_1(t)=\lambda_1(t,\tau)$$

variază și ea în raport cu acest parametru.
Ne propunem să determinăm valoarea funcției caracteristice $F_h(n,t)$ pentru

$$n=\frac{a}{\tau},$$

unde a este o constantă.
Din formula (X.6,4) reiese că este necesar să se evalueze asimptotic

$$[\lambda_1(t,\tau)]^{\frac{a}{\tau}}.$$

Se iese din domeniul distribuției normale numai dacă

$$\lambda_1(t,0)=1.$$

Cum $\lambda_1(t,\tau)$ este rădăcina ecuației

$$\begin{vmatrix} p_{11}(\tau)\,e^{ti}-\lambda & p_{12}(\tau) & \dots & p_{1m}(\tau) \\ p_{12}(\tau)\,e^{ti} & p_{22}(\tau)-\lambda & \dots & p_{2m}(\tau) \\ \dots & \dots & \dots & \dots \\ p_{m1}(\tau)\,e^{ti} & p_{m2}(\tau) & \dots & p_{mm}(\tau)-\lambda \end{vmatrix}=0, \qquad \text{(X.6,5)}$$

pentru ca să avem

$$\lambda_1(t,0)=1,$$

se vede imediat că trebuie să avem, cu notațiile de mai sus,

$$G_{11}(1,0)=0. \qquad \text{(X.6,6)}$$

Am arătat că, dacă unul dintre minorii principali ai matricei $D(\lambda,t,0)$ este nul, matricea este decompozabilă, adică printr-o permutare identică a liniilor și coloanelor poate fi adusă la forma

$$\begin{vmatrix} P & Q \\ R & S \end{vmatrix},$$

unde P şi S sînt matrice pătrate, iar R este egal cu zero. Prin urmare, pentru

$$\tau=0$$

matricea stochastică trebuie să fie

$$\begin{vmatrix} p_{11}(0) & p_{12}(0) \dots p_{1h}(0) & p_{1,h+1}(0) & p_{1,h+2}(0) & \dots p_{1m}(0) \\ p_{21}(0) & p_{22}(0) \dots p_{2h}(0) & p_{2,h+1}(0) & p_{2,h+2}(0) & \dots p_{2m}(0) \\ \dots & \dots & \dots & \dots & \dots \\ p_{h1}(0) & p_{h2}(0) \dots p_{hh}(0) & p_{h,h+1}(0) & p_{h,h+2}(0) & \dots p_{hm}(0) \\ 0 & 0 \dots 0 & p_{h+1,h+1}(0) & p_{h+1,h+2}(0) & \dots p_{h+1\,m}(0) \\ 0 & 0 \dots 0 & p_{h+2,h+1}(0) & p_{h+2,h+2}(0) & \dots p_{h+2,m}(0) \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 \dots 0 & p_{m,h+1}(0) & p_{m,h+2}(0) & \dots p_{mm}(0) \end{vmatrix} . \tag{X.6,7}$$

Să examinăm mai întîi cazul

$$h=1,$$

care revine la a presupune că matricea $G_{11}(1, 0)$ este indecompozabilă. Insemnăm prin P_h minorul corespunzător elementului $p_{hk}-1$ în determinantul $D(1, 0, 0)$ şi prin B_h minorul corespunzător aceluiaşi element în determinantul $G_{11}(1, 0, 0)$.

Admitem că derivatele de ordinul întîi ale funcţiilor $p_{h1}(\tau)$ există şi că avem $p'_{h1}(0)=a_h$.

Din (X.6,5) rezultă că

$$\lim_{\tau\to 0}[\lambda_1(t, \tau)]^{\frac{a_s}{\tau}} = e^{-a\frac{\sum a_s B_s}{\sum B_s}\cdot\frac{1-e^{ti}}{1-\left(1+\frac{\sum P_s}{\sum B_s}\right)e^{ti}}} . \tag{X.6,8}$$

Trecînd la limită şi pentru celelalte elemente din formula (X.6,4), ajungem la următorul rezultat

$$F_1(n, t, \tau) \simeq \frac{1-p_{11}}{1-p_{11}e^{ti}}\, e^{-a\frac{\sum a_s P_s}{\sum P_s}\cdot\frac{1-e^{ti}}{1-p_{11}e^{ti}}}, \tag{X.6,9}$$

$$F_j(n,\ t,\ \tau)\simeq e^{-a\frac{\sum a_s P_s}{\sum P_s}\cdot\frac{1-e^{ti}}{1-p_{11}e^{ti}}} \qquad \left(n=\frac{a}{\tau}\right). \qquad \text{(X.6,10)}$$

Ele extind schema lui Poisson pentru lanțurile lui Markov.

Pentru cazul unui lanț care ia numai două valori, formulele (X.6,9) și (X,6,10) au fost găsite de B. D. Koopman. Este interesant de observat că ele rămîn neschimbate cînd se consideră un lanț care ia un număr finit de valori.

Să trecem acum la cazul

$$h>1.$$

Plecînd de la matricea stochastică (X.6,7), să însemnăm prin Δ determinantul

$$\begin{vmatrix} p_{11}(0)-1 & p_{12}(0) & \dots & p_{1h}(0) \\ p_{21}(0) & p_{22}(0)-1 & \dots & p_{2h}(0) \\ \dots & \dots & \dots & \dots \\ p_{h1}(0) & p_{h2}(0) & \dots & p_{hh}(0)-1 \end{vmatrix}$$

și prin Δ_1 minorul corespunzător elementului $p_{11}(0)-1$.

Avem, corespunzător formulei (X.6,8),

$$\lim_{\tau\to 0}[\lambda_1(t,\ \tau)]^{\frac{a}{\tau}}=e^{-ak\frac{1-e^{ti}}{1-\left(1+\frac{\Delta}{\Delta_1}\right)ti}},$$

unde k este o constantă.

Rezultă atunci că avem

$$F_1(n,\ t,\ \tau)\simeq\frac{\Delta}{(\Delta+\Delta_1)e^{ti}-\Delta_1}e^{-ak\frac{1-e^{ti}}{1-\left(1+\frac{\Delta_1}{\Delta}\right)e^{ti}}} \qquad \left(n=\frac{a}{\tau}\right),$$

$$F_j(n,\ t,\ \tau)\simeq\frac{e^{ti}\left[\sum\limits_{k=h+1}^{m}G_k^k+\sum\limits_{k=h+1}^{m}G_{1j}^{1k}\right]-\sum\limits_{k=h+1}^{m}G_{1j}^{1k}}{e^{ti}\left[\sum\limits_{k=h+1}^{m}G_k^k+\sum\limits_{k=h+1}^{m}G_{ik}^{ik}\right]-\sum\limits_{k=h+1}^{m}G_{ik}^{ik}}e^{-ak\frac{1-e^{ti}}{1-\left(1+\frac{\Delta_1}{\Delta}\right)e^{ti}}},$$

$$\left(j=2,\ 3,\dots,h;\ n=\frac{a}{\tau}\right)$$

$$F_j(n,\ t,\ \tau)=e^{-ak\frac{1-e^{ti}}{1-\left(1+\frac{\Delta_1}{\Delta}\right)e^{ti}}} \qquad \left(n=\frac{a}{\tau},\ j=h+1,\ h+2,\ldots,m\right),$$

unde am însemnat prin
$F_j(n,\ t,\ \tau)$ funcţia caracteristică în cazul cînd probabilitatea egalităţii $x_1=a_k$ este p_{ik} $(k=1,\ 2,\ \ldots,\ n)$;

G_h^{λ} determinantul obţinut din $D(1,\ 0,\ 0)$ prin suprimarea coloanei de rangul λ şi a liniei de rangul h;
G_{ij}^{kl} determinantul obţinut din $D(1,\ 0,\ 0)$ prin suprimarea coloanelor de rangul i, j şi a liniilor de rangul k şi l.
In modul acesta se pot da şi alte generalizări ale schemei lui Poisson, care prezintă nu numai interes teoretic, prin faptul că extinde pentru variabilele dependente rezultate clasice din domeniul variabilelor independente, dar şi interes practic prin aplicaţiile lor.

§ 7. Momentele distribuţiei Poisson multidimensionale

Pentru determinarea momentelor unei distribuţii Poisson cu $m-1$ dimensiuni, plecăm de la funcţia caracteristică a acestei distribuţii, a cărei expresie, după cum am văzut, este asimptotic egală cu

$$\varphi(u_1,\ u_2,\ldots,u_{m-1},\ \lambda_1,\ \lambda_2,\ldots,\lambda_{m-1})\simeq e^{\sum_{k=1}^{m-1}\lambda_k(e^{u_k i}-1)}. \qquad (X.7,1)$$

Această expresie a funcţiei caracteristice corespunde unei variabile vectoriale aleatorii, astfel ca vectorii $e_1(1,\ 0,\ldots,\ 0)$, $e_2(0,\ 1,\ 0,\ldots,\ 0)$,..., $e_{m-1}(0,\ldots,0,\ 1,\ 0)$ să apară respectiv de α_1 ori, α_2 ori,..., α_{m-1} ori în n probe consecutive independente.
Aplicînd logaritmii în expresia (X.7,1), avem

$$\ln\varphi=\sum_{k=1}^{m-1}\lambda_k(e^{u_k i}-1). \qquad (X.7,2)$$

Derivînd relaţia (X.7,2) succesiv în raport cu $u_1,\ u_2,\ \ldots,\ u_{m-1}$, vom obţine momentele distribuţiei, după ce mai înainte vom face $u_1=u_2=\ldots=u_{m-1}=0$ şi vom observa că

$$[\varphi(u_1,\ u_2,\ldots,u_{m-1};\ \lambda_1,\ \lambda_2,\ldots,\lambda_{m-1})]_{u_1=u_2=\ldots=u_{m-1}}=1.$$

Scanned with OKEN Scanner

Astfel

$$M(\alpha_h) = \frac{1}{i}\left[\frac{\partial\varphi}{\partial u_h}\right]_{u_1=u_2=\ldots=u_{m-1}=0} = \lambda_h,$$

$$M(\alpha_h^2) = \frac{1}{i^2}\left[\frac{\partial^2\varphi}{\partial u_h^2}\right]_{u_1=u_2=\ldots=u_{m-1}=0} = \lambda_h + \lambda_h^2,$$

$$M(\alpha_h\alpha_k) = \frac{1}{i^2}\left[\frac{\partial^2\varphi}{\partial u_h\,\partial u_k}\right]_{u_1=u_2=\ldots=u_{m-1}=0} = \lambda_h\lambda_k,$$

$$M(\alpha_h^3) = \frac{1}{i^3}\left[\frac{\partial^3\varphi}{\partial u_h^3}\right]_{u_1=u_2=\ldots=u_{m-1}=0} = \lambda_h + 3\lambda_h^2 + \lambda_h^3,$$

$$M(\alpha_h^2\alpha_k) = \frac{1}{i^3}\left[\frac{\partial^3\varphi}{\partial u_h^2\,\partial u_k}\right]_{u_1=u_2=\ldots=u_{m-1}=0} = \lambda_h\lambda_k(1+\lambda_h),$$

$$M(\alpha_h\alpha_k\alpha_l) = \frac{1}{i^3}\left[\frac{\partial^3\varphi}{\partial u_h\,\partial u_k\,\partial u_l}\right]_{u_1=u_2=\ldots=u_{m-1}=0} = \lambda_h\lambda_k\lambda_l$$

$$(h \neq k \neq l = 1,\ 2,\ldots,\ m-1),$$

. . . . . .

Deducem imediat momentele centrate, ținînd seama de observația pe care am făcut-o cînd am determinat probabilitatea $P_n(\alpha_1, \alpha_2,\ldots,\alpha_m)$ ca, în n probe independente, vectorii $e_1(1, 0,\ldots,0)$, $e_2(0, 1, 0,\ldots,0)$,... ..., $e_m(0,\ldots, 0, 1)$ să apară respectiv de α_1 ori, α_2 ori,..., α_m ori. Am stabilit atunci că această probabilitate este, pentru valori mari ale lui n, asimptotic egală cu

$$P_n(\alpha_1,\ \alpha_2,\ldots,\alpha_m) \simeq \prod_{k=1}^{m-1} \frac{\lambda_k^{\alpha_k}}{\alpha_k!} e^{-\lambda_k}.$$

De unde, observația importantă că $\alpha_1,\ \alpha_2,\ldots,\alpha_{m-1}$ sînt independente între ele și deci $\alpha_1-\lambda_1,\ \alpha_2-\lambda_2,\ldots,\alpha_{m-1}-\lambda_{m-1}$ sînt, de asemenea, independente între ele ($\lambda_1,\ \lambda_2,\ldots,\lambda_{m-1}$ fiind niște constante).
In acest caz deducem

$$M(\mu_h) = M(\alpha_h-\lambda_h) = 0,$$

$$M(\mu_h^2) = M(\alpha_h-\lambda_h)^2 = M(\alpha_h^2)-\lambda_h^2 = \lambda_h,$$

$$M(\mu_h \mu_k) = M[(\alpha_h - \lambda_h)(\alpha_k - \lambda_k)] = M(\alpha_h - \lambda_h) M(\alpha_k - \lambda_k) = 0,$$
$$M(\mu_h^3) = M(\alpha_h - \lambda_h)^3 = M(\alpha_h)^3 - 3\lambda_h^2 M(\alpha_h) + 2\lambda_h^3 = \lambda_h,$$
$$M(\mu_h^2 \mu_k) = M[(\alpha_h - \lambda_h)^2 (\alpha_k - \lambda_k)] = M(\alpha_h - \lambda_h)^2 M(\alpha_k - \lambda_k) = 0,$$
$$M(\mu_h \mu_k \mu_l) = M[(\alpha_h - \lambda_h)(\alpha_k - \lambda_k)(\alpha_l - \lambda_l)] = 0$$
$$(h \neq k \neq l = 1, 2, \ldots, m-1).$$

§ 8. Distribuţia binomială generalizată a lui Poisson

Despre o generalizare a distribuţiei binomiale am mai avut ocazia să vorbim, atunci cînd ne-am ocupat de distribuţia multinomială. Am văzut că această distribuţie este dată de suma de variabile aleatorii independente

$$X_1, X_2, \ldots, X_n,$$

unde fiecare dintre aceste variabile poate lua valorile

$$a_1, a_2, \ldots, a_m$$

cu probabilităţile respective

$$p_1, p_2, \ldots, p_m \qquad \left(\sum_{i=1}^{m} p_i = 1\right).$$

Cu alte cuvinte, în fiecare dintre cele n probe independente, avem distribuţia

$$X_i \begin{pmatrix} a_1, & a_2, \ldots, & a_m \\ p_1, & p_2, \ldots, & p_m \end{pmatrix} \qquad (i = 1, 2, \ldots, n).$$

Dar, o generalizare a distribuţiei binomiale mai putem obţine, considerînd succesiunea de variabile aleatorii independente

$$X_1, X_2, \ldots, X_n,$$

fiecare dintre variabilele X_h $(h = 1, 2, \ldots, n)$ putînd lua valorile unu sau zero cu probabilităţile respective p_h şi q_h $(p_h + q_h = 1)$.
In acest caz, fiecare probă dă naştere distribuţiei

$$X_h \begin{pmatrix} 1 & 0 \\ p_h & q_h \end{pmatrix} \qquad (h = 1, 2, \ldots, n).$$

Distribuţia binomială generalizată a lui Poisson este dată de variabila

$$Y_n = X_1 + X_2 + \ldots + X_n.$$

Scanned with OKEN Scanner

Asemenea distribuţii se întîlnesc în analiza unor experienţe, atunci cînd există o probabilitate constantă de apariţie a unui eveniment într-un interval de timp scurt. Pe această perioadă de timp, putem asimila fenomenul din punct de vedere probabilistic cu o schemă bernoulliană.

Pentru o perioadă de timp mai mare, vom avea deci o succesiune de scheme Bernoulli cu probabilităţi diferite $(p_1, q_1), (p_2, q_2), \ldots, (p_n, q_n)$, adică tocmai generalizarea lui Poisson a distribuţiei binomiale.

Momentele acestei distribuţii se pot obţine în mod simplu, folosind teoremele generale privind momentele sumelor de variabile independente

$$M_1 = M(Y_n) = M\left(\sum_{h=1}^{n} X_h\right) = \sum_{h=1}^{n} M(X_h) = \sum_{h=1}^{n} p_h,$$

$$M_2 = M(Y_n^2) = M\left(\sum_{h=1}^{n} X_h\right)^2 = \sum_{h=1}^{n} M(X_h^2) + 2\sum_{h\neq k} M(X_h)\,M(X_k) =$$

$$= \sum_{h=1}^{n} p_h + 2\sum_{h\neq k} p_h p_k,$$

$$M_3 = M(Y_n^3) = M\left(\sum_{h=1}^{n} X_h\right) = \sum_{h=1}^{n} M(X_h^3) + 3\sum_{h\neq k} M(X_h^2)\,M(X_k) +$$

$$+ 6\sum_{h\neq k\neq l} M(X_h)\,M(X_k)\,M(X_l) =$$

$$= \sum_{h=1}^{n} p_h + 3\sum_{h\neq k} p_h p_k + 6\sum_{h\neq k\neq l} p_h p_k p_l,$$

$$M_4 = M(Y_n^4) = M\left(\sum_{h=1}^{n} X_h\right)^4 = \sum_{h=1}^{n} M(X_h^4) + 4\sum_{h\neq k} M(X_h^3)\,M(X_k) +$$

$$+ 6\sum_{h\neq k} M(X_h^2)\,M(X_k^2) + 12\sum_{h\neq k\neq l} M(X_h^2)\,M(X_k)\,M(X_l) +$$

$$+ 24\sum_{h\neq k\neq l\neq r} M(X_h)\,M(X_k)\,M(X_l)\,M(X_r) =$$

$$= \sum_{h=1}^{n} p_h + 10\sum_{h\neq k} p_h p_k + 12\sum_{h\neq k\neq l} p_h p_k p_l + 24\sum_{h\neq k\neq l\neq r} p_h p_k p_l p_r,$$

. . . . . .

Asemenea distribuţii se întîlnesc în analiza unor experienţe, atunci cînd există o probabilitate constantă de apariţie a unui eveniment într-un interval de timp scurt. Pe această perioadă de timp, putem asimila fenomenul din punct de vedere probabilistic cu o schemă bernoulliană.

Pentru o perioadă de timp mai mare, vom avea deci o succesiune de scheme Bernoulli cu probabilităţi diferite $(p_1, q_1), (p_2, q_2), \ldots, (p_n, q_n)$, adică tocmai generalizarea lui Poisson a distribuţiei binomiale.

Momentele acestei distribuţii se pot obţine în mod simplu, folosind teoremele generale privind momentele sumelor de variabile independente

$$M_1 = M(Y_n) = M\left(\sum_{h=1}^{n} X_h\right) = \sum_{h=1}^{n} M(X_h) = \sum_{h=1}^{n} p_h,$$

$$M_2 = M(Y_n^2) = M\left(\sum_{h=1}^{n} X_h\right)^2 = \sum_{h=1}^{n} M(X_h^2) + 2 \sum_{h \neq k} M(X_h)\, M(X_k) =$$

$$= \sum_{h=1}^{n} p_h + 2 \sum_{h \neq k} p_h p_k,$$

$$M_3 = M(Y_n^3) = M\left(\sum_{h=1}^{n} X_h\right) = \sum_{h=1}^{n} M(X_h^3) + 3 \sum_{h \neq k} M(X_h^2)\, M(X_k) +$$

$$+ 6 \sum_{h \neq k \neq l} M(X_h)\, M(X_k)\, M(X_l) =$$

$$= \sum_{h=1}^{n} p_h + 3 \sum_{h \neq k} p_h p_k + 6 \sum_{h \neq k \neq l} p_h p_k p_l,$$

$$M_4 = M(Y_n^4) = M\left(\sum_{h=1}^{n} X_h\right)^4 = \sum_{h=1}^{n} M(X_h^4) + 4 \sum_{h \neq k} M(X_h^3)\, M(X_k) +$$

$$+ 6 \sum_{h \neq k} M(X_h^2)\, M(X_k^2) + 12 \sum_{h \neq k \neq l} M(X_h^2)\, M(X_k)\, M(X_l) +$$

$$+ 24 \sum_{h \neq k \neq l \neq r} M(X_h)\, M(X_k)\, M(X_l)\, M(X_r) =$$

$$= \sum_{h=1}^{n} p_h + 10 \sum_{h \neq k} p_h p_k + 12 \sum_{h \neq k \neq l} p_h p_k p_l + 24 \sum_{h \neq k \neq l \neq r} p_h p_k p_l p_r,$$

. . . . . .

De asemenea, obţinem pentru momentele centrate

$$m_1 = M(Y_n - M_1) = 0,$$

$$m_2 = M(Y_n - M_1)^2 = M_2 - M_1^2 = \sum_{h=1}^{n} p_h +$$

$$+ 2 \sum_{h \neq k} p_h p_k - \left(\sum_{h=1}^{n} p_h \right)^2 = \sum_{h=1}^{n} p_h q_h.$$

Dispersia $m_2 = \sigma^2$ a variabilei Y_n este egală cu suma dispersiilor fiecărei variabile X_h $(h = 1, 2, \ldots, n)$.
Funcţia caracteristică este evident

$$\varphi(t, n) = \prod_{h=1}^{n} (p_h e^{ti} + q_h).$$

Derivînd succesiv în raport cu t funcţia caracteristică şi făcînd după aceea $t = 0$, vom deduce momentele distribuţiei binomiale generalizate a lui Poisson, pe care le-am găsit pe cale directă.

§ 9. Dispersia subnormală

Distribuţia binomială generalizată a lui Poisson este dată, după cum am arătat în paragraful precedent, de variabila $\sum_{h=1}^{n} X_h$, unde fiecare dintre variabilele X_h dă naştere la distribuţia

$$X_h \begin{pmatrix} 1, & 0 \\ p_h, & q_h \end{pmatrix} \qquad (p_h + q_h = 1).$$

Să presupunem că în locul schemei binomiale generalizate a lui Poisson se dă succesiunea de variabile aleatorii independente

$$Z_1, Z_2, \ldots, Z_n,$$

fiecare din ele dînd naştere la distribuţia

$$Z_h \begin{pmatrix} 1, & 0 \\ p, & q \end{pmatrix} \qquad (p + q = 1),$$

unde am notat $p = \dfrac{p_1 + p_2 + \ldots + p_n}{n}$.

Ne propunem să determinăm diferența între dispersia variabilei $\sum_{h=1}^{n} X_h$ și a variabilei $\sum_{h=1}^{n} Z_h$.

Dispersia variabilei aleatorii $\sum_{h=1}^{n} X_h$ a fost calculată mai înainte și am găsit că este egală cu

$$\sum_{h=1}^{n} p_h q_h.$$

Dispersia variabilei aleatorii $\sum_{h=1}^{n} Z_h$ este evident npq, deoarece sîntem în cazul distribuției binomiale simple cu probabilități constante în fiecare probă.
Să considerăm, acum, expresiile

$$\sum_{h=1}^{n} (p_h - p)^2 = \sum_{h=1}^{n} p_h^2 - np^2,$$
$$\sum_{h=1}^{n} (q_h - q)^2 = \sum_{h=1}^{n} q_h^2 - nq^2. \qquad \text{(X.9,1)}$$

Adunînd egalitățile (X.9,1) și observînd că

$$\sum_{h=1}^{n} (p_h - p)^2 = \sum_{h=1}^{n} (q_h - q)^2,$$

obținem

$$2\sum_{h=1}^{n} (p_h - p)^2 = \sum_{h=1}^{n} p_h^2 + \sum_{h=1}^{n} q_h^2 - np^2 - nq^2 =$$
$$= \sum_{h=1}^{n} (p_h + q_h)^2 - 2\sum_{h=1}^{n} p_h q_h - n(p+q)^2 + 2npq,$$

de unde rezultă

$$npq = \sum_{h=1}^{n} p_h q_h + \sum_{h=1}^{n} (p_h - p)^2, \qquad \text{(X.9,2)}$$

deoarece

$$\sum_{h=1}^{n} (p_h + q_h)^2 = n(p+q)^2 = n.$$

Relaţia (X.9,2) ne arată că dispersia distribuţiei binomiale dată de variabila $\sum_{h=1}^{n} Z_h$ este mai mare decît dispersia distribuţiei binomiale generalizate a lui Poisson dată de variabila $\sum_{h=1}^{n} X_h$, întrucît diferenţa dintre cele două dispersii este

$$\sum_{h=1}^{n} (p_h - p)^2 > 0.$$

In mod concret, avem următoarea schemă.
Fie n urne $(u_1, u_2, \ldots, u_n)$, fiecare dintre ele conţinînd A_h bile albe şi $N - A_h$ bile negre cu probabilitatea de a extrage o bilă albă egală cu

$$p_h = \frac{A_h}{N} \qquad (h = 1, 2, \ldots, n).$$

Dacă facem cîte o extracţie din fiecare dintre cele n urne, avem schema probabilistică a distribuţiei binomiale generalizate a lui Poisson şi atunci valoarea medie a numărului de bile albe va fi

$$M\left(\sum_{h=1}^{n} X_h\right) = \sum_{h=1}^{n} p_h = \frac{\sum_{h=1}^{n} A_h}{N}. \qquad \text{(X.9,3)}$$

In locul celor n urne să luăm acum o singură urnă U, care să conţină un număr $\sum_{h=1}^{n} A_h$ de bile albe şi $n \cdot N - \sum_{h=1}^{n} A_h$ bile negre, adică toate bilele cuprinse în cele n urne U_i $(i = 1, 2, \ldots, n)$.
Probabilitatea de a scoate o bilă albă din această urnă este

$$p = \frac{\sum_{h=1}^{n} A_h}{nN} = \frac{\sum_{h=1}^{n} p_h}{n}.$$

Rezultă, prin urmare, că dacă facem n extracţii succesive din urna U, vom obţine în medie un număr de bile albe egal cu

$$M\left(\sum_{h=1}^{n} Z_h\right) = np = \frac{\sum_{h=1}^{n} A_h}{N}, \qquad \text{(X.9,4)}$$

deoarece sîntem în cazul schemei probabilistice a distribuţiei binomiale cu probabilităţi constante de a se produce evenimentul în fiecare dintre cele n probe.

Relaţiile (X.9,3) şi (X.9,4) exprimă faptul că, atît în primul cît şi în al doilea caz, obţinem în medie acelaşi număr de bile albe, iar dispersia, în baza relaţiei (X.9,2) este mai mare în al doilea caz, decît în primul.

Se spune, după o nomenclatură datorită lui W. Lexis, care a studiat cel dintîi verificarea unor anumite ipoteze statistice cu ajutorul dispersiei, că schema distribuţiei binomiale generalizate a lui Poisson, potrivit proprietăţii de mai sus, dă naştere la o *dispersie subnormală.*

După cum se observă, existenţa dispersiei subnormale arată că probabilitatea variază de la probă la probă.

§ 10. Dispersia supranormală

Să considerăm funcţia de frecvenţă, dată de expresia

$$\frac{1}{m}\sum_{h=1}^{m} P_n(\alpha_h) \quad (m>1), \tag{X.10,1}$$

unde $P_n(\alpha_h)$ este probabilitatea ca în n probe, evenimentul E să se producă de α_h ori, în baza schemei probabilistice a lui J. Bernoulli. Se vede uşor că expresia (X.10,1) este o funcţie de frecvenţă, deoarece avem

$$\sum_{\alpha_h=0}^{n}\left(\frac{1}{m}\sum_{h=1}^{m} P_n(\alpha_h)\right)=\frac{1}{m}\sum_{h=1}^{m}\left(\sum_{\alpha_h=0}^{n} P_n(\alpha_h)\right)=\frac{1}{m}\cdot m=1.$$

Ne propunem să calculăm dispersia distribuţiei a cărei funcţie de frecvenţă este expresia (X.10,1).

Ţinînd seama de momentele distribuţiei binomiale, rezultă

$$M_1=\sum_{\alpha_h=0}^{n}\alpha_h\left(\frac{1}{m}\sum_{h=1}^{m} P_n(\alpha_h)\right)=\frac{1}{m}\sum_{h=1}^{m}\left(\sum_{\alpha_h=0}^{n}\alpha_h\cdot P_n(\alpha_h)\right)=\frac{1}{m}\sum_{h=1}^{m} np_h=np,$$

dacă notăm $\frac{\sum_{h=1}^{m} p_h}{m} = p$,

$$M_2 = \sum_{\alpha_h=0}^{n} \alpha_h^2 \left(\frac{1}{m} \sum_{h=1}^{m} P_n(\alpha_h) \right) = \frac{1}{m} \sum_{h=1}^{m} \left(\sum_{\alpha_h=0}^{n} \alpha_h^2 P_n(\alpha_h) \right) =$$

$$= \frac{1}{m} \sum_{h=1}^{m} [np_h + n(n-1)p_h^2] = np + \frac{n(n-1)}{m} \sum_{h=1}^{m} p_h^2,$$

de unde deducem

$$m_2 = M_2 - M_1^2 = np + \frac{n(n-1)}{m} \sum_{h=1}^{m} p_h^2 - n^2p^2 = npq +$$

$$+ \frac{n(n-1)}{m} \sum_{h=1}^{m} p_h^2 - n(n-1)p^2 = npq + n(n-1) \left(\frac{\sum_{h=1}^{m} p_h^2}{m} - p^2 \right) =$$

$$= npq + n(n-1) \frac{\sum_{h=1}^{m} (p_h - p)^2}{m}. \qquad \text{(X.10,2)}$$

Din relaţia (X.10,2) observăm că dispersia distribuţiei noastre este mai mare decît dispersia distribuţiei binomiale dată de variabila $\sum_{h=1}^{m} Z_h$, fiecare variabilă Z_h dînd naştere la distribuţia

$$Z_h \begin{pmatrix} 1 & 0 \\ p & q \end{pmatrix} \qquad (p+q=1).$$

De data aceasta se spune că schema distribuţiei considerate, în baza proprietăţii de mai sus, dă naştere la o *dispersie supranormală*.

Egalitatea dintre cele două dispersii va avea loc, evident, numai dacă $p_h = p$, ceea ce este totuna, dacă probabilităţile $p_h (h = 1, 2, \ldots, m)$ vor fi egale între ele, adică în cazul distribuţiei binomiale.

Această egalitate va avea loc de asemenea şi în cazul dispersiei subnormale, despre care am vorbit în paragraful precedent.

Scanned with OKEN Scanner

In mod concret, vom putea reprezenta şi exemplul următor.
Se dau urnele $u_h (h=1, 2, \ldots, m)$, unde fiecare dintre ele conţine A_h bile albe şi $N-A_h$ bile negre, cu probabilitatea de a scoate o bilă albă egală cu

$$p_h=\frac{A_h}{N}.$$

Se fac cîte n extracţii succesive din fiecare urnă u_h.
Atunci, valoarea medie a numărului de bile albe din n extracţii succesive va fi media valorilor medii de bile albe ieşite din fiecare urnă u_h

$$M_1=\frac{1}{m}\sum_{h=1}^{m} np_h. \qquad \text{(X.10,3)}$$

Să presupunem acum că, în locul celor m urne, am avea o singură urnă U, care conţine $\sum_{h=1}^{m} A_h$ bile albe şi $mN-\sum_{h=1}^{m} A_h$ bile negre.
Probabilitatea de a scoate o bilă albă din această urnă va fi egală cu

$$p=\frac{\sum_{h=1}^{m} A_h}{mN}=\frac{\sum_{h=1}^{m} p_h}{m}.$$

Deoarece sîntem în cazul unei distribuţii binomiale cu aceleaşi probabilităţi în fiecare probă, valoarea medie a numărului de bile albe în urma a n extracţii succesive va fi dată de

$$M_1=np=\frac{1}{m}\sum_{h=1}^{m} np_h. \qquad \text{(X.10,4)}$$

Cele două expresii (X.10,3) şi (X.10,4) ale lui M_1 ne confirmă că obţinem aceeaşi valoare medie de bile albe în n extracţii succesive, atît în primul, cît şi în al doilea caz.
In ceea ce priveşte dispersia, după cum rezultă din (X.10,2), ea este mai mare în primul caz, decît în al doilea.

Scanned with OKEN Scanner

§ 11. Distribuţia binomială cu exponent negativ

Fie E un eveniment de probabilitate constantă de a se produce p şi de probabilitate contrară $q=p$. Vom supune evenimentul E la α probe pînă ce el se realizează de m ori, m fiind un număr întreg pozitiv, dat. Numărul de probe α pe care trebuie să le facem pentru ca să obţinem de m ori evenimentul E este o variabilă aleatorie şi evident $\alpha \geqslant m$.

Vom nota prin $P_m(\alpha)$ probabilitatea ca evenimentul E să se verifice de m ori în cele α probe, prin A o realizare a evenimentului E şi prin B, o nerealizare a lui în fiecare din aceste probe.

Vom avea, astfel, o succesiune de forma

$$AA \ldots BA \ldots BBA, \qquad \text{(X.11,1)}$$

unde A va intra de m ori, iar B de $(\alpha - m)$ ori.

De la început să observăm că, în această succesiune, A trebuie să fie ultimul, pentru că dacă ar fi B ultimul, ar însemna că evenimentul E s-ar fi realizat de m ori în primele $\alpha - 1$ probe.

Teoremă. *Probabilitatea ca, în α probe, evenimentul E să se verifice de m ori, m fiind un număr întreg şi pozitiv dat $(\alpha \geqslant m)$ este egală cu*

$$P_m(\alpha) = C_{\alpha-1}^{m-1} p^m q^{\alpha-m}. \qquad \text{(X.11,2)}$$

Probabilitatea de a avea loc o succesiune de forma (X.11,1) este evident $p^m \cdot q^{\alpha-m}$.

Atunci, pentru determinarea probabilităţii de a avea toate succesiunile posibile cu A la urmă, va trebui să asociem această realizare a evenimentului E în ultima probă α cu numărul succesiunilor care se pot obţine în primele $\alpha - 1$ probe, în care evenimentul E s-a verificat de $m-1$ ori.

Numărul acestor succesiuni va fi $C_{\alpha-1}^{m-1}$, iar probabilitatea căutată se va scrie

$$P_m(\alpha) = C_{\alpha-1}^{m-1} p^m \cdot q^{\alpha-m},$$

Deci, teorema este demonstrată.

Rezultă, ca o consecinţă imediată a teoremei, că

$$\sum_{\alpha=m}^{\infty} P_m(\alpha) = 1.$$

In adevăr, din (X 11,2) deducem

$$\sum_{\alpha=m}^{\infty} P_m(\alpha) = p^m \sum_{\alpha=m}^{\infty} C_{\alpha-1}^{m-1} q^{\alpha-m} = p^m \sum_{s=0}^{\infty} C_{m+s-1}^{m-1} q^s = p^m (1-q)^{-m} = 1,$$

unde am notat $\alpha - m = s$.

Observăm că probabilitatea $P_m(\alpha)$ este termenul general al dezvoltării binomului $p^m(1-q)^{-m} = \left(\frac{1}{p} - \frac{q}{p}\right)^{-m}$ cu exponent negativ, de unde provine numele distribuției dată de variabila α, *de distribuție binomială cu exponent negativ.*

Aplicațiile distribuției binomiale cu exponent negativ se întîlnesc în domeniul contagiunii.

Să presupunem că o populație microbiană este formată dintr-un număr N de membri. Asupra lor repetăm un același tratament, care supune microbii la riscul de a fi loviți de un medicament.

Notînd cu p probabilitatea ca un microb să fie atins într-un tratament și presupunînd că un același microb moare dacă este atins în m tratamente, atunci se vede ușor, în baza celor arătate mai sus, că distribuția statistică a microbilor se desfășoară după o distribuție binomială cu exponent negativ.

În sfîrșit, aceste distribuții intervin, de asemenea, în teoria estimației, despre care vom vorbi mai departe.

§ 12. Momentele distribuției binomiale cu exponent negativ

Studiul distribuției binomiale cu exponent negativ se face mai ușor considerînd întîi cazul $m = 1$.

Atunci, ținînd seama de (X.11,2) vom avea

$$P_1(\alpha) = C_{\alpha-1}^{0} p q^{\alpha-1} = p q^{\alpha-1},$$

de unde, rezultă

$$M_1 = M(\alpha) = \sum_{\alpha=1}^{\infty} \alpha p q^{\alpha-1} = p(1 + 2q + 3q^2 + \ldots) = \frac{p}{(1-q)^2} = \frac{1}{p}.$$

Pentru determinarea momentelor de ordin superior lui 1, vom calcula mai înainte momentele factoriale.

Funcția generatoare a momentelor factoriale se va scrie

$$\Phi(t) = \sum_{\alpha=1}^{\infty} t^{\alpha} p q^{\alpha-1} = pt \sum_{\alpha=1}^{\infty} (qt)^{\alpha-1} = \frac{pt}{1-qt} = -\frac{p}{q} + \frac{p}{q}(1-qt)^{-1}.$$

Prin urmare deducem

$$M_{(1)}=\left(\frac{d\Phi}{dt}\right)_{t=1}=[p\,(1-qt)^{-2}]_{t=1}=\frac{1}{p},$$

$$M_{(2)}=\left(\frac{d^2\Phi}{dt^2}\right)_{t=1}=[1\cdot 2\cdot pq\,(1-qt)^{-3}]_{t=1}=\frac{2!q}{p^2},$$

$$M_{(3)}=\left(\frac{d^3\Phi}{dt^3}\right)_{t=1}=[1\cdot 2\cdot 3\cdot pq^2(1-qt)^{-4}]_{t=1}=\frac{3!\,q^2}{p^3},$$

. . . . . .

$$M_{(r)}=\left(\frac{d^r\Phi}{dt^r}\right)_{t=1}=[1\cdot 2\ldots rpq^{r-1}(1-qt)^{-(r+1)}]_{t=1}=\frac{r!\,q^{r-1}}{p^r}.$$

De la momentele factoriale se poate trece uşor la celelalte momente ale distribuţiei, cunoscînd că momentul M_r este o funcţie liniară de momentele factoriale $M_{(k)}$ $(k=1,\ 2,\ldots,\ r)$

$$M_r=T_r^1M_{(1)}+T_r^2M_{(2)}+\ldots+T_r^rM_{(r)},$$

unde T_r^k sînt numerele lui Stirling de speţa a doua.
Avem deci

$$M_1=M_{(1)}=\frac{1}{p},$$

$$M_2=M_{(1)}+M_{(2)}=\frac{1}{p}+\frac{2q}{p^2},\qquad (X.12,1)$$

$$M_3=M_{(1)}+3M_{(2)}+M_{(3)}=\frac{1}{p}+\frac{6q}{p^2}+\frac{6q}{p^3},$$

$$M_4=M_{(1)}+7M_{(2)}+6M_{(3)}+M_{(4)}=\frac{1}{p}+\frac{14q}{p^2}+\frac{36q^2}{p^3}+\frac{24q^3}{p^3},$$

. . . . . .

In sfirşit, obţinem

$$m_1=0,$$

$$m_2=M_2-M_1^2=\frac{q}{p^2},$$

. . . . . .

Am calculat momentele distribuției dată de variabila α, care reprezintă numărul probelor pînă ce evenimentul E s-a realizat o singură dată, adică pentru cazul $m=1$.

Rămîne acum să rezolvăm problema mai generală de a determina momentele distribuției binomiale cu exponent negativ pentru cazul în care m este oarecare.

Vom presupune că am făcut m probe și anume, α_1 probe pînă ce evenimentul E s-a realizat o dată, α_2 probe pînă ce evenimentul E_2 s-a mai realizat o dată etc.

Variabilele aleatorii $\alpha_1, \alpha_2, \ldots, \alpha_m$ sînt independente între ele și fiecare urmează, evident, o lege binomială cu exponent negativ ($m=1$).

Considerînd variabila aleatorie

$$\alpha' = \alpha_1 + \alpha_2 + \ldots + \alpha_m .$$

care reprezintă numărul de probe în care s-a verificat evenimentul E de m ori, vom putea obține ușor momentele căutate, calculînd momentele distribuției variabilei α'.

Astfel, vom găsi

$$M_1 = M(\alpha') = M\left(\sum_{i=1}^{m} \alpha_i\right) = \sum_{i=1}^{m} M(\alpha_i) = \frac{m}{p},$$

$$M_2 = M(\alpha'^2) = M\left(\sum_{i=1}^{m} \alpha_i\right)^2 = \sum_{i=1}^{m} M(\alpha_i^2) + 2\sum_{i \neq j} M(\alpha_i)\, M(\alpha_j) =$$

$$= \frac{m^2}{p^2} + \frac{mq}{p^2}, \qquad \text{(X.12,2)}$$

. . . . . .

$$m_1 = 0,$$

$$m_2 = M_2 - M_1^2 = \frac{mq}{p^2},$$

. . . . . .

§ 13. Funcția caracteristică a distribuției binomiale cu exponent negativ

Funcția caracteristică a variabilei α_k, care ne dă numărul probelor în care evenimentul E s-a realizat o singură dată, este, în baza definiției sale,

$$\varphi_1(t) = \sum_{s=0}^{\infty} pq^s e^{(s+1)ti} = pe^{ti} \sum_{s=0}^{\infty} (qe^{ti})^s = \frac{pe^{ti}}{1 - qe^{ti}}, \qquad \text{(X.13,1)}$$

unde am notat $s = \alpha_k - 1$.

Aplicînd teorema funcţiei caracteristice a sumei de variabile independente, putem să scriem funcţia caracteristică a variabilei $\alpha' = \sum_{k=1}^{m} \alpha_k$

$$\varphi_m(t) = \left(\frac{pe^{ti}}{1-qe^{ti}}\right)^m. \qquad (X.13,2)$$

Dacă derivăm în raport cu t expresiile (X.13,1) şi (X.13,2) ale funcţiilor caracteristice şi vom face apoi pe $t=0$, vom regăsi aceleaşi momente date de relaţiile (X.12,1), respectiv (X.12,2), după formula

$$M_r = \frac{1}{i}\left(\frac{d^r\varphi}{dt^r}\right)_{t=0}.$$

Teoremă. *Dacă avem o lege binomială cu exponent negativ şi facem pe m, care reprezintă frecvenţa evenimentului E, să crească nemărginit, atunci această lege tinde către legea normală a lui Laplace.*

Demonstraţia teoremei se face folosind funcţia caracteristică a variabilei reduse $\frac{p\left(\alpha - \frac{m}{p}\right)}{\sqrt{mq}}$, care se scrie

$$\Phi_m(t) = M\left[e^{\frac{p\left(\alpha-\frac{m}{p}\right)}{\sqrt{mq}}ti}\right] = e^{-\sqrt{\frac{m}{q}}ti} M\left(e^{\alpha\frac{pt}{\sqrt{mq}}i}\right) =$$

$$= e^{-\sqrt{\frac{m}{q}}ti}\varphi_m\left(\frac{pt}{\sqrt{mq}}\right)$$

sau, ţinînd seama de (X.13,2),

$$\Phi_m(t) = e^{-\sqrt{\frac{m}{q}}ti}\left(\frac{pe^{\frac{pti}{\sqrt{mq}}}}{1-qe^{\frac{pti}{\sqrt{mq}}}}\right)^m = \left(\frac{pe^{-\sqrt{\frac{q}{m}}ti}}{1-qe^{\frac{pti}{\sqrt{mq}}}}\right)^m =$$

$$= \left(\frac{1}{p}e^{\frac{\sqrt{q}}{\sqrt{m}}ti} - \frac{q}{p}e^{\frac{ti}{\sqrt{mq}}}\right)^{-m}. \qquad (X.13,3)$$

Aplicînd cunoscuta proprietate de dezvoltare în serie a lui e^x, funcția caracteristică $\Phi_m(t)$ se mai scrie

$$\Phi_m(t)=\left[\frac{1}{p}\left(1+\frac{\sqrt{q}\,ti}{\sqrt{m}}-\frac{qt^2}{2m^2}+\theta\frac{q^{\frac{2}{3}}t^3}{6m^{\frac{3}{2}}}\right)-\frac{q}{p}\left(1+\frac{ti}{\sqrt{mq}}-\frac{t^2}{2mq}+\right.\right.$$

$$\left.\left.+\theta\frac{3}{6(mq)^{\frac{3}{2}}}\right)\right]=\left[1+\frac{t^2}{2m}+O\left(m^{-\frac{3}{2}}\right)\right]^{-m},$$

unde $|\theta|<1$, iar $O\left(m^{-\frac{3}{2}}\right)$ este o funcție asimptotică de ordinul lui $m^{-\frac{3}{2}}$.

Insă, din (X.13,3) rezultă

$$\lim_{m\to\infty}\Phi_m(t)=\lim_{m\to\infty}\left[1+\frac{t^2}{2m}+O\left(m^{-\frac{3}{2}}\right)\right]^{-m}=e^{-\frac{t^2}{2}}. \quad \text{(X.13,4)}$$

Relația de limită (X.13,4) ne arată că șirul de funcții caracteristice $\Phi_m(t)$ tinde către funcția caracteristică $e^{-\frac{t^2}{2}}$ a legii normale $N(0, 1)$ cînd m crește la infinit.

Atunci, funcțiile de repartiție corespunzătoare vor tinde, de asemenea, către funcția de repartiție a legii normale $N(0, 1)$ și teorema este, astfel, demonstrată.

§ 14. Interpretarea geometrică a distribuției binomiale și a distribuției binomiale cu exponent negativ

Interpretarea geometrică a distribuției binomiale și a distribuției binomiale cu exponent negativ rezultă foarte ușor dintr-o reprezentare grafică a numărului de probe în care evenimentul E s-a realizat și în care el nu s-a realizat.

Să luăm în sistemul de coordonate rectangulare punctul $M(\alpha, \beta)$, a cărui abscisă α reprezintă numărul de cîte ori s-a realizat evenimentul E, iar ordonata β, numărul de cîte ori nu s-a realizat acest eveniment.

Este clar că cele două coordonate α şi β nu pot fi decît numere întregi şi pozitive sau nule.

Prin urmare, în planul XOY determinat de cele două axe de coordonate, fiecare punct $M(\alpha, \beta)$ va corespunde la $\alpha+\beta$ probe, în care evenimentul E s-a verificat de α ori şi nu s-a verificat de β ori (fig. X.14,1).

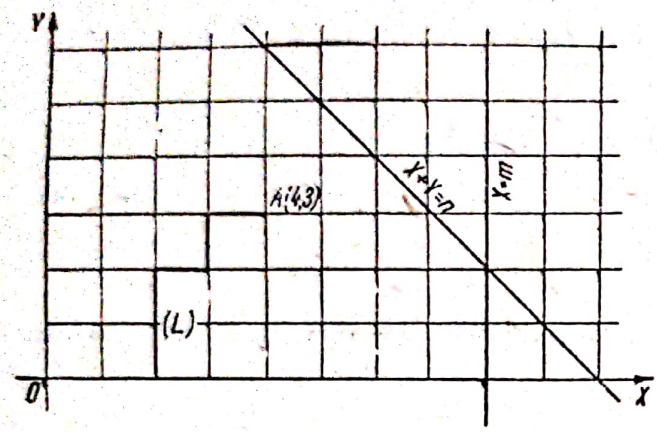


Fig. X.14,1

Să trasăm un drum (L) care porneşte, de exemplu, de la origine la un punct $A(4, 3)$. Drumul (L) este caracterizat prin faptul că merge sau la dreapta sau în sus.

Punctul A corespunde, deci, la şapte probe, în care evenimentul E s-a verificat de patru ori şi nu s-a verificat de trei ori.

Drumul (L) ne arată, însă, şi modul cum sau succedat realizările evenimentului E în cursul celor şapte probe.

Astfel, în primele două probe, evenimentul E s-a realizat de două ori, în următoarele două probe nu s-a realizat de loc, în proba a cincia s-a realizat, în a şasea nu s-a realizat şi în a şaptea iarăşi s-a realizat.

Să considerăm dreapta

$$(D) \qquad X+Y=n \qquad (n>0)$$

pe care sînt aşezate punctele de coordonate

$$(0, n), \quad (1, n-1), \ldots, (\alpha, n-\alpha), \ldots, (n, 0).$$

Probabilitatea ca un drum (L) să atingă dreapta (D) în punctul de coordonate $(\alpha, n-\alpha)$ este $C_n^\alpha p^\alpha q^{n-\alpha}$, deoarece punctul de coordonate $(\alpha, n-\alpha)$ corespunde la n probe în care evenimentul E de probabilitate p şi q $(p+q=1)$ s-a verificat de α ori şi nu s-a verificat de $n-\alpha$ ori.

Rezultă, deci, că dreapta (D) va corespunde distribuţiei binomiale.

Să considerăm, de asemenea, dreapta

$$(D') \qquad X=m \qquad (m>0)$$

pe care se găsesc punctele de coordonate

$$(m,\ 0),\quad (m,\ 1),\ \ldots\ ,(m,\ k),\ \ldots\ ,(m,\ m),\ \ldots$$

Probabilitatea ca drumul (L) să atingă dreapta (D') în punctul de coordonate $(m,\ k)$ este $C_{m+k-1}^{m-1}\ p^m q^k = C_{\alpha-1}^{m-1}\ p^m \cdot q^{\alpha-m}$, unde am notat $m+k=\alpha$.

In adevăr, punctul de coordonate $(m,\ k)$ corespunde la o succesiune de $\alpha=m+k$ probe, în care s-a realizat evenimentul E de m ori şi nu s-a realizat de k ori.

In acest caz, dreapta (D') va corespunde distribuției binomiale cu exponent negativ.

Este evident că, dacă vom considera dreapta

$$Y=m \qquad (m>0),$$

ea va corespunde la fel unei distribuții binomiale cu exponent negativ.

Capitolul XI

Alte distribuții clasice

§ 1. Curbele de distribuție ale lui Pearson

Să considerăm ecuația diferențială

$$\frac{dy}{dx} = \frac{(x-c)\,y}{c_0+c_1x+c_2x^2}, \qquad \text{(XI.1,1)}$$

unde c_0, c_1, c_2 şi c sînt nişte constante.

Se vede uşor că această relație diferențială generalizează ecuația diferențială

$$\frac{dy}{dx} = -\frac{(x-m)\,y}{\sigma^2}, \qquad \text{(XI.1,2)}$$

pe care o verifică funcția $y=\frac{1}{\sqrt{2\pi}\cdot\sigma}\,e^{-\frac{(x-m)^2}{2\sigma^2}}$, adică, funcția de frecvență a distribuției normale $N(m,\ \sigma)$.

In adevăr, dacă facem în (XI.1,1)

$$c_0=-\sigma^2, \quad c_1=c_2=0, \quad c=m,$$

obţinem ecuaţia diferenţială (XI.1,2).

Pearson şi-a propus să determine funcţiile de frecvenţă care verifică ecuaţia diferenţială (XI.1,1).

Este evident că nu orice soluţie a ecuaţiei diferenţiale (XI.1,1) poate să fie o funcţie de frecvenţă, deoarece o funcţie de frecvenţă trebuie să fie întotdeauna pozitivă. De aceea, vom lua soluţiile acelei ecuaţii diferenţiale numai pe intervalele în care ele sînt pozitive, iar pe intervalele în care ele sînt negative, le vom înlocui cu valoarea zero.

Din cauza formei particulare a acestei diferenţiale (XI.1,1) soluţia ei generală este de forma

$$f(x,\ \alpha)=\alpha g(x),$$

unde α este constanta arbitrară de integrare, care se poate determina din condiţia

$$\alpha\int_H g(x)\,dx=1,$$

H fiind mulţimea intervalelor pentru care avem $\alpha\cdot g(x)\geqq 0$.

In modul acesta, soluţia generală $\alpha\cdot g(x)$ a ecuaţiei diferenţiale (XI.1,1) dă naştere funcţiei de distribuţie

$$F(x,\ \alpha)=\alpha\int_{-\infty}^{x} g(x)\,dx \qquad (x\in H),$$

$$F(x,\ \alpha)=0 \qquad (x\in\overline{H}).$$

$\overline{E}$ fiid complementara lui E.

Vom presupune îndeplinită condiţia

$$x^m g(x)\to 0, \qquad \text{(XI.1,3)}$$

care are loc pentru orice distribuţie finită (m întreg şi pozitiv).

Curbele lui Pearson, definite de ecuaţia diferenţială (XI.1,3), sînt folosite pentru ajustarea distribuţiilor statistice empirice.

In acest scop, să ne imaginăm că ne este dată o distribuţie statistică ale cărei momente sînt

$$M_0,\ M_1,\ M_2,\ \ldots,\ M_n,\ \ldots$$

Scanned with OKEN Scanner

Atunci, vom determina constantele c_0, c_1, c_2 şi c, pentru ca primele patru momente ale distribuţiei Pearson să coincidă cu primele patru momente ale distribuţiei date.

Din (XI.1,1) deducem

$$(c_0+c_1x+c_2x^2)\,dy=(x-c)\,ydx. \qquad \text{(XI.1,4)}$$

Inmulţind ecuaţia diferenţială (XI.1,4) cu x^{n-1} şi integrînd între limitele $-\infty$ şi $+\infty$, obţinem

$$\int_{-\infty}^{+\infty}(c_0+c_1x+c_2x^2)\,x^{n-1}\,dy=\int_{-\infty}^{+\infty}(x-c)\,x^{n-1}\,ydx. \qquad \text{(XI.1,5)}$$

Integrînd prin părţi membrul întîi din (XI.1,5) avem

$$\int_{-\infty}^{+\infty}(c_0+c_1x+c_2x^2)\,x^{n-1}dy=[(c_0+c_1x+c_2x^2)\,x^{n-1}\,y]_{-\infty}^{+\infty}=$$

$$=\int_{-\infty}^{+\infty}[c_0(n-1)\,x^{n-2}+c_1nx^{n-1}+c_2(n+1)\,x^n]\,ydx=$$

$$=-\int_{-\infty}^{+\infty}[c_0(n-1)\,x^{n-2}+c_1nx^{n-2}+c_2(n+1)\,x^n]\,ydx,$$

deoarece, în baza condiţiei (XI.1,3), avem

$$[(c_0+c_1x+c_2x^2)\,x^{n-1}\cdot y]_{-\infty}^{+\infty}=0.$$

Egalitatea (XI.1,5) se poate scrie acum

$$-c_0(n-1)\,M_{n-2}-c_1nM_{n-1}-c_2(n+1)\,M_n=M_n-cM_{n-1}$$

sau

$$[1+c_2(n+1)]\,M_n+(c_1n-c)\,M_{n-1}+c_0(n-1)\,M_{n-2}=0. \qquad \text{(XI.1,6)}$$

Observăm că (XI.1,6) este o relaţie de recurenţă între trei momente consecutive, care are loc pentru $n=2, 3, \ldots$

Dacă $n=1$, relaţia de recurenţă se reduce la

$$(1+2c_2)\,M_1+(c_1-c)\,M_0=0 \qquad (M_0=1).$$

Pentru determinarea constantelor c_0, c_1, c_2 şi c vom face $n=1, 2, 3, 4$ în relaţia (XI.1,6) şi vom obţine astfel următorul sistem de ecuaţii

$$\begin{aligned}
&(1+2c_2)\,M_1+c_1-c=0,\\
&(1+3c_2)\,M_2+(2c_1-c)\,M_1+c_0=0,\\
&(1+4c_2)\,M_3+(3c_1-c)\,M_2+2c_0M_1=0,\\
&(1+5c_2)\,M_4+(4c_1-c)\,M_3+3c_0M_2=0.
\end{aligned} \tag{XI.1,7}$$

Rezolvînd sistemul de ecuaţii (XI.1,7) şi alegînd originea axelor de coordonate astfel ca să coincidă cu media distribuţiei date, găsim următoarele valori pentru constantele c_0, c_1, c_2 şi c

$$\begin{aligned}
c_0&=-\frac{m_2(4m_2m_4-3m_3^2)}{10m_2m_4-18m_2^3-12m_3^2},\\
c=c_1&=-\frac{m_3(m_4+3m_2^2)}{10m_2m_4-18m_2^3-12m_3^2},\\
c_2&=-\frac{2m_2m_4-6m_2^3-3m_3^2}{10m_2m_4-18m_2^3-12m_3^2}.
\end{aligned} \tag{XI.1,8}$$

Momentele M_k le-am înlocuit prin momentele m_k, deoarece prin mutarea originii în punctul de abscisă egală cu M_1, rezultă

$$M_1=0, \quad M_k=m_k \qquad (k\geqq 2).$$

Introducînd dispersia, oblicitatea şi excesul, determinate prin relaţiile cunoscute

$$\sigma^2=m_2, \quad \gamma_1=\frac{m_3}{m_2^{\frac{3}{2}}}, \quad \gamma_2=\frac{m_4}{m_2^2}=3,$$

soluţiile (XI.1,8) devin

$$c_0=-\frac{\sigma^2(4\gamma_2-3\gamma_1^2+12)}{2(5\gamma_2-6\gamma_1^2+6)}=-\sigma^2\cdot\frac{r+1}{r-2},$$

$$c=c_1=-\frac{\sigma\gamma_1(\gamma_2+6)}{2(5\gamma_2-6\gamma_1+6)}=-\frac{\sigma\gamma_1}{2}\cdot\frac{r+2}{r-2},$$

$$c_2=-\frac{2\gamma_2-3\gamma_1^2}{2(5\gamma_2-6\gamma_1+6)}=\frac{1}{r-2},$$

unde

$$r = \frac{6(\gamma_2 - \gamma_1^2 + 2)}{3\gamma_1^2 - 2\gamma_2}.$$

Problema ajustării este acum rezolvată, deoarece prin determinarea constantelor c_0, c_1, c_2 și c și a constantei arbitrare de integrare α, care, după cum am văzut, se determină din ecuația $\alpha \int_H g(x)\,dx = 1$, va rezulta o singură curbă de distribuție, corespunzătoare distribuției considerate.

Dacă am dori să obținem o ajustare mai bună, vom putea folosi o ecuație diferențială de forma

$$\frac{dy}{dx} = \frac{(x-c)\,y}{c_0 + c_1 x + c_2 x^2 + \ldots + c_p x^p}. \qquad \text{(XI.1,9)}$$

In acest caz, se înțelege că va fie nevoie de un număr mai mare de momente, însă rezolvarea ecuației diferențiale (XI.1,9) va fi cu atît mai grea, cu cît p este mai mare.

§ 2. Tipuri de curbe Pearson

Curbele Pearson se împart în mai multe tipuri, după forma soluției generale a ecuației diferențiale (XI.1,1)

$$\frac{dy}{dx} = \frac{(x-c)\,y}{c_0 + c_1 x + c_2 x^2}.$$

Tipul 1. Dacă trinomul $c_0 + c_1 x + c_2 x^2$ are două rădăcini reale, $\lambda_1 \neq \lambda_2$, soluția generală a ecuației diferențiale de mai sus se poate scrie sub forma

$$y = \alpha (x-\lambda_1)^{m_1} \cdot (\lambda_2 - x)^{m_2} \qquad (\lambda_1 < x < \lambda_2). \qquad \text{(XI.2,1)}$$

Curbele a căror funcție de frecvență este dată de (XI.2,1) reprezintă *curbele lui Pearson de tipul* 1.

Constanta α se va determina din condiția

$$\alpha \int_{\lambda_1}^{\lambda_2} (x-\lambda_1)^{m_1} \cdot (\lambda_2 - x)^{m_2}\,dx = 1.$$

In particular, dacă $\lambda_1 = 0$ şi $\lambda_2 = 1$, curba de distribuţie (XI.2,1) devine

$$y = \frac{1}{B(m_1+1,\ m_2+1)} \cdot x^{m_1}(1-x)^{m_2},$$

unde

$$B(m_1+1,\ m_2+1) = \int_0^1 x^{m_1}(1-x)^{m_2}\,dx.$$

Tipul 2. In cazul $\lambda_1 = -\lambda_2 = \lambda$, funcţia de frecvenţă (XI.2,1) ia forma

$$y = C(\lambda^2 - x^2)^m \qquad (-\lambda < x < \lambda),\ \left(m > -\frac{1}{2}\right). \qquad \text{(XI.2,2)}$$

Curbele particulare de tipul 1 a căror funcţie de frecvenţă este dată de (XI.2,2) alcătuiesc *curbele lui Pearson de tipul* 2.

Tipul 3. Dacă constanta $c_2 = 0$, atunci soluţia generală a ecuaţiei diferenţiale (XI.1,1) va fi

$$y = \alpha(c_0 + c_1 x)^m e^{-ax};\quad \left(x > -\frac{c_0}{c_1}\right),\ (a>0),\ \left(m > -\frac{1}{2}\right), \qquad \text{(XI.2,3)}$$

unde constanta α se determină prin relaţia

$$\alpha \int_{-\frac{c_0}{c_1}}^{\infty} (c_0 + c_1 x)^m e^{-ax}\,dx = 1.$$

Curbele cu funcţia de frecvenţă (XI.2,3) corespund *curbelor lui Pearson de tipul* 3.

In cazul particular cînd $c_0 = 0$, $c_1 = 1$, $m = \frac{n}{2} - 1$ şi $a = \frac{1}{2}$, funcţia de frecvenţă (XI.2,3) se reduce la

$$y = \alpha x^{\frac{n}{2}-1} \cdot e^{-\frac{x^2}{2}}.$$

Distribuţia cu această funcţie de frecvenţă se numeşte *distribuţia* χ^2 *a lui Helmert-Pearson*, după numele lui Helmert, care a găsit-o prima oară în 1876 şi a lui Pearson care a regăsit-o cu 30 ani în urmă. Despre distribuţia lui Helmert-Pearson ne vom ocupa în paragraful următor.

Scanned with OKEN Scanner

Tipul 4. Dacă rădăcinile trinomului $c_0+c_1x+c_2x^2$ sînt complexe, atunci soluţia generală a ecuaţiei diferenţiale (XI.1,1) este

$$y=\alpha(c_0+c_1x+c_2x^2)^m e^{n\,\mathrm{arctg}\,\frac{c_1+2c_2x}{\sqrt{4c_0c_2-c_1^2}}}, \qquad \text{(XI.2,4)}$$

constanta α determinîndu-se din relaţia $\int\limits_{-\infty}^{+\infty} y\,dx=1$.

Curbele lui Pearson de tipul 4 sînt acelea a căror funcţie de frecvenţă este dată de expresia (XI.2,4).

Tipul 5. Trinomul $c_0+c_1x+c_2x^2$ poate avea o rădăcină dublă $\lambda_1=\lambda_2=\lambda$, adică $c_1^2-4c_0c_2=0$, $c_2\neq 0$.

Atunci, soluţia generală a ecuaţiei diferenţiale (XI.1,1) se va scrie

$$y=\alpha(x-\lambda)^m\cdot e^{-\frac{n}{x-\lambda}} \qquad \text{(XI.2,5)}$$

$$(x>\lambda),\quad (m>-1),\quad (n>0).$$

Constanta α se va determina acum din ecuaţia

$$\alpha\int\limits_{\lambda}^{\infty}(x-\lambda)^m e^{-\frac{n}{x-\lambda}}\cdot dx=1.$$

Curbele a căror funcţie de frecvenţă este de forma (XI.2,5) definesc *curbele lui Pearson de tipul* 5.

Tipul 6. *Curbele lui Pearson de tipul* 6 sînt, ca şi cele de tipul 2, un caz particular al curbelor lui Pearson de tipul 1.

Funcţia de frecvenţă este dată de soluţia generală

$$y=\alpha(x-\lambda_1)^{m_1}\cdot(x-\lambda_2)^{m_2}$$

$$(\lambda_1<x;\quad \lambda_1<\lambda_2)\quad (m_2>-1)\quad (m_1+m_2<-1)$$

cu α determinat din condiţia

$$\alpha\int\limits_{-\infty}^{+\infty}(x-\lambda_1)^{m_1}\cdot(x-\lambda_2)^{m_2}\,dx=1.$$

Tipul 7. *Curbele lui Pearson de tipul* 7 sînt cazuri particulare ale curbelor lui Pearson de tipul 4. Ele sînt remarcabile prin faptul că, la rîndul lor, conţin ca un caz particular distribuţia normală.

Dacă

$$c_0=1, \quad c_1=0, \quad c_2=1, \quad m=-\frac{p+s}{2} \quad \text{şi} \quad n=0,$$

obţinem *distribuţia lui Student*, despre care vom vorbi mai departe.

Ca o observaţie generală constatăm că, dacă unii dintre parametrii m_1, m_2, m, n tind către infinit, atunci obţinem distribuţia normală în multe dintre legile lui Pearson.

Curbele lui Pearson, după cum se vede, se deduc din ecuaţia diferenţială (XI.1,1) fără nici o justificare teoretică.

In primul rînd, ele au o semnificaţie practică, deoarece conţin drept cazuri particulare cele mai importante distribuţii întrebuinţate în aplicaţiile statisticii matematice.

In al doilea rînd, cu ajutorul formulelor analitice care le exprimă, se poate face o ajustare bună a distribuţiilor statistice care rezultă din experienţe.

Curbele lui Pearson pot căpăta însă o justificare teoretică.

Astfel, s-au construit sisteme de urne, care dau naştere la procese aleatorii ale căror distribuţii sînt curbe Pearson. Importanţa acestor scheme rezidă în faptul că ele pot fi luate ca model pentru multe fenomene de masă ca de pildă în cazul contagiunii.

In modul acesta, curbele lui Pearson au căpătat în ultima vreme şi suportul teoretic care le lipsea.

§ 3. Distribuţia χ^2 a lui Helmert-Pearson

Această distribuţie, după cum am arătat mai înainte, a fost descoperită de Helmert în 1876 şi pusă în valoare cu 30 ani mai tîrziu de către Pearson.

Considerînd şirul de variabile aleatorii independente

$$x_1, x_2, \ldots, x_s,$$

unde fiecare dintre aceste variabile urmează legea normală $N(0, \sigma)$, ne propunem să găsim funcţia de frecvenţă a distribuţiei, dată de variabila

$$y=x_1^2+x_2^2+\ldots+x_s^2.$$

1) Vom lua variabilele $x_1, x_2, \ldots, x_s$ drept coordonatele unui punct $M(x_1, x_2, \ldots, x_n)$ dintr-un spaţiu euclidian cu s dimensiuni.

Legea de probabilitate a unei variabile x_i fiind

$$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x_i^2}{2\sigma^2}},$$

probabilitatea elementară în punctul $M(x_1, x_2, \ldots, x_s)$, adică probabilitatea ca acest punct să cadă în intervalul

$$(x_1, x_1+dx_1\,;\; x_2, x_2+dx_2\,;\ldots;\, x_s, x_s+dx_s)$$

va fi

$$\frac{1}{(2\pi)^{\frac{s}{2}}\sigma^s} e^{-\frac{\sum_{i=1}^{s} x_i^2}{2\sigma^2}} dx_1 \cdot dx_2 \ldots dx_s$$

şi deci, funcţia de repartiţie se va scrie

$$F(y) = P\left(\sum_{i=1}^{s} x_i^2 \leqslant y\right) =$$

$$= \frac{1}{(2\pi)^{\frac{s}{2}}\sigma^s} \int\int \ldots \int_{(S)} e^{-\frac{\sum_{i=1}^{s} x_i^2}{2\sigma^2}} dx_1 \cdot dx_2 \ldots dx_s, \qquad \text{(XI.3,1)}$$

unde domeniul de integrare (S) este volumul hipersferei $x_1^2+x_2^2+\ldots$ $\ldots+x_s^2=\rho$ cu centrul în origine şi raza egală cu $\rho^{\frac{1}{2}}$.

Integrala multiplă (XI.3,1) se va calcula uşor, dacă vom face următoarea schimbare de variabile în coordonate polare

$$\begin{aligned}
x_1 &= \rho^{\frac{1}{2}} \cos\theta_1 \cos\theta_2 \ldots \cos\theta_{s-1},\\
x_2 &= \rho^{\frac{1}{2}} \cos\theta_1 \cos\theta_2 \ldots \cos\theta_{s-2} \sin\theta_{s-1},\\
x_3 &= \rho^{\frac{1}{2}} \cos\theta_1 \cos\theta_2 \ldots \cos\theta_{s-3} \sin\theta_{s-2},\\
&\cdots\cdots\cdots\cdots\cdots\cdots\\
x_s &= \rho^{\frac{1}{2}} \sin\theta_1.
\end{aligned} \qquad \text{(XI.3,2)}$$

Se constată fără greutate că determinantul funcţional al transformării de variabile (XI.3,2) este

$$\frac{D(x_1, x_2, \ldots, x_s)}{D(\rho, \theta_1, \theta_2, \ldots, \theta_{s-1})} = \frac{1}{2}\rho^{\frac{s-2}{2}} \cos^{s-2}\theta_1 \cos^{s-3}\theta_2 \ldots \cos\theta_{s-2}.$$

Atunci, funcţia de repartiţie (XI.3,1) se va scrie

$$F(y) = \frac{1}{(2\pi)^{\frac{s}{2}}\sigma^s} \int\int \ldots \int e^{-\frac{\rho}{2\sigma^2}} \frac{1}{2} \left| \rho^{\frac{s-2}{2}} \cos^{s-2}\theta_1 \cos^{s-3}\theta_2 \ldots \right.$$

$$\left. \ldots \cos\theta_{s-2} \right| d\rho\, d\theta_1 \ldots d\theta_{s-1}, \qquad \text{(XI.3,3)}$$

unde

$$0 \leq \rho \leq y, \quad 0 \leq \theta_k \leq 2\pi \quad (k = 1, 2, \ldots, s-1).$$

In acest caz, expresia (XI.3,3) a funcţiei de repartiţie se poate pune sub forma

$$F(y) =$$

$$= \frac{1}{2(2\pi)^{\frac{s}{2}}\sigma^s} \int e^{-\frac{\rho}{2\sigma^2}} \rho^{\frac{s-2}{2}} d\rho \int_0^{2\pi} \left| \cos^{s-2}\theta_1 \right| d\theta_1 \int_0^{2\pi} \left| \cos^{s-3}\theta_2 \right| d\theta_2 \ldots \int_0^{2\pi} d\theta_{s-1} =$$

$$= K \int_0^y e^{-\frac{\rho}{2\sigma^2}} \rho^{\frac{s-2}{2}} d\rho, \qquad \text{(XI.3,4)}$$

unde am notat

$$K = \frac{1}{2(2\pi)^{\frac{s}{2}}\sigma^s} \int_0^{2\pi} \left| \cos^{s-2}\theta_1 \right| d\theta_1 \int_0^{2\pi} \left| \cos^{s-3}\theta_2 \right| d\theta_2 \ldots \int_0^{2\pi} d\theta_{s-1}.$$

Din egalitatea

$$K \int_0^\infty e^{-\frac{\rho}{2\sigma^2}} \cdot \rho^{\frac{s-2}{2}} \cdot d\rho = 1,$$

deducem

$$K = \frac{1}{\int_0^\infty e^{-\frac{\rho^2}{2\sigma^2}} \cdot \rho^{\frac{s-2}{2}} \cdot d\rho}. \qquad \text{(XI.3,5)}$$

Scanned with OKEN Scanner

Efectuînd substituţia $x=\frac{\rho}{2\sigma^2}$ şi ţinînd seama de definiţia funcţiei $\Gamma(x)$, relaţia (XI.3,5) devine

$$K=\frac{1}{2^{\frac{s}{2}}\sigma^s\int_0^{\infty}x^{\frac{s}{2}-1}\cdot e^{-x}\cdot dx}=\frac{1}{2^{\frac{s}{2}}\cdot\sigma^s\cdot\Gamma\left(\frac{s}{2}\right)}. \qquad \text{(XI.3,6)}$$

Introducînd, acum, valoarea constantei K, dată de relaţia (XI.3,6) în expresia (XI.3,4) a funcţiei de repartiţie, obţinem

$$F(y)=\frac{1}{2^{\frac{s}{2}}\cdot\sigma^s\cdot\Gamma\left(\frac{s}{2}\right)}\int_0^{y}\rho^{\frac{s-2}{2}}\cdot e^{-\frac{\rho}{2\sigma^2}}\cdot d\rho. \qquad \text{(XI.3,7)}$$

In sfîrşit, derivînd expresia (XI.3,7) de mai sus în raport cu y, rezultă funcţia de frecvenţă a distribuţiei χ^2 a lui Helmert-Pearson

$$f(y)=\frac{dF(y)}{dy}=\frac{1}{2^{\frac{s}{2}}\cdot\sigma^2\Gamma\left(\frac{s}{2}\right)}y^{\frac{s}{2}-1}\cdot e^{-\frac{y}{2\sigma^2}}$$

sau, schimbînd variabila y cu x,

$$f(x)=\frac{1}{2^{\frac{s}{2}}\cdot\sigma^s\cdot\Gamma\left(\frac{s}{2}\right)}x^{\frac{s}{2}-1}\cdot e^{-\frac{x}{2\sigma^2}}. \qquad \text{(XI.3,8)}$$

2) La expresia (XI.3,8) a funcţiei de frecvenţă $f(x)$ se poate ajunge, de asemenea, determinînd legea de probabilitate a variabilei x_i^2. Pentru aceasta, vom stabili mai întîi funcţia de repartiţie a acestei variabile.
Vom scrie deci

$$F(y)=P(x_i^2\leqslant y)=P(-\sqrt{y}\leqslant x_i\leqslant\sqrt{y})=$$

$$=\frac{1}{\sqrt{2\pi}\cdot\sigma}\int_{-\sqrt{y}}^{\sqrt{y}}e^{-\frac{x_i^2}{2\sigma^2}}dx_i=\frac{2}{\sqrt{2\pi}\cdot\sigma}\int_0^{\sqrt{y}}e^{-\frac{x_i^2}{2\sigma^2}}dx_i,$$

deoarece, fiecare dintre variabilele x_i $(i=1, 2, \ldots, s)$ urmează legea normală $N(0, \sigma)$, de unde, rezultă

$$f(y)=\frac{dF(y)}{dy}=\frac{1}{\sqrt{2\pi}\cdot\sigma\cdot y^{\frac{1}{2}}}e^{-\frac{y}{2\sigma^2}},$$

adică legea de probabilitate a variabilei x_i^2.
Acum, vom putea găsi funcţia caracteristică a variabilei x_i^2, care, în baza definiţiei sale, se scrie

$$M(e^{yti})=\frac{1}{\sqrt{2\pi}\cdot\sigma}\int_0^\infty y^{-\frac{1}{2}}e^{\frac{y}{2\sigma^2}}e^{yti}dy=$$

$$=\frac{1}{\sqrt{2\pi}\sigma}\int_0^\infty y^{\frac{s}{2}-1}e^{-\frac{1-2\sigma^2ti}{2\sigma^2}y}dy. \qquad \text{(XI.3,9)}$$

Din definiţia funcţiei $\Gamma(x)$ se ştie că

$$\int_0^\infty x^{p-1}\cdot e^{-ax}\cdot dx=\frac{\Gamma(p)}{a^p}. \qquad \text{(XI.3,10)}$$

Atunci, expresia (XI.3,9) a funcţiei caracteristice, observînd că $p=\frac{1}{2}$ şi $a=\frac{1-2\sigma^2ti}{2\sigma^2}$, devine

$$M(e^{yti})=\frac{1}{\sqrt{2\pi}\cdot\sigma}\cdot\frac{\Gamma\left(\frac{1}{2}\right)}{\left(\frac{1-2\sigma^2ti}{2\sigma^2}\right)^{\frac{1}{2}}}=\frac{1}{(1-2\sigma^2ti)^{\frac{1}{2}}}=(1-2\sigma^2ti)^{-\frac{1}{2}}$$

$$\left[\Gamma\left(\frac{1}{2}\right)=\sqrt{\pi}\right].$$

Aplicînd teorema funcţiei caracteristice a sumei de variabile independente, obţinem imediat funcţia caracteristică a variabilei $\sum_{i=1}^{s}x_i^2$

$$\varphi(t)=(1-2\sigma^2ti)^{-\frac{s}{2}}. \qquad \text{(XI.3,11)}$$

Scanned with OKEN Scanner

Din (XI.3,10) deducem

$$\frac{a^p}{\Gamma(p)}\int_0^\infty x^{p-1}\cdot e^{-ax}dx=1,$$

ceea ce înseamnă că funcţia

$$f(x)=\frac{a^p}{\Gamma(p)}\cdot x^{p-1}e^{-ax} \qquad \text{(XI.3,12)}$$

este funcţie de frecvenţă a distribuţiei care conduce la funcţia $\Gamma(p)$. Funcţia caracteristică, corespunzătoare acestei distribuţii va fi

$$\psi(t)=\frac{a^p}{\Gamma(p)}\int_0^\infty x^{p-1}e^{-ax}e^{x\,ti}dx=\frac{a^p}{\Gamma(p)}\int_0^\infty x^{p-1}e^{-(a-ti)\,x}dx=$$

$$=\frac{a^p}{\Gamma(p)}\cdot\frac{\Gamma(p)}{(a-ti)\,p}=\left(1-\frac{1}{a}ti\right)^{-p}. \qquad \text{(XI.3,13)}$$

Din compararea funcţiilor caracteristice $\varphi(t)$ şi $\psi(t)$ date respectiv de expresiile (XI.3,11) şi (XI.3,13), care corespund amîndouă la o distribuţie a funcţiei gama, obţinem

$$p=\frac{s}{2} \quad \text{şi} \quad a=\frac{1}{2\sigma^2}.$$

Inlocuind valorile p şi a în (XI,3,12), regăsim funcţia de frecvenţă (XI.3,8) a distribuţiei χ^2 a lui Helmert-Pearson

$$f(x)=\frac{\left(\frac{1}{2\sigma^2}\right)^{\frac{s}{2}}}{\Gamma\left(\frac{s}{2}\right)}x^{\frac{s}{2}-1}e^{-\frac{x}{2\sigma^2}}=\frac{1}{2^{\frac{s}{2}}\sigma^s\Gamma\left(\frac{s}{2}\right)}x^{\frac{s}{2}-1}e^{-\frac{x}{2\sigma^2}}.$$

3. Funcţia de frecvenţă a distribuţiei χ^2 a lui Helmert-Pearson se mai poate stabili pe baza unor consideraţii geometrice.

Să presupunem, astfel, că, într-un spaţiu euclidian cu s dimensiuni, se dă hipersfera cu centrul în origine şi cu raza $R=y^{\frac{1}{2}}$

$$x_1^2+x_2^2+\ldots+x_s^2=y.$$

Funcţia de frecvenţă $f(y)\,dy$ va fi, evident, egală cu probabilitatea ca un punct $M(x_1, x_2, \ldots, x_s)$ din acest spaţiu cu s dimensiuni să cadă în domeniul (V), cuprins între două hipersfere cu centrele în origine şi cu razele respectiv egale cu $y^{\frac{1}{2}}$ şi $(y+dy)^{\frac{1}{2}}$.

Cunoscînd că probabilitatea elementară pentru punctul $M(x_1, x_2, \ldots, x_s)$ este

$$\frac{1}{(2\pi)^{\frac{s}{2}}\sigma^s} e^{-\frac{\sum_{i=1}^{s} x_i^2}{2\sigma^2}} dx_1\, dx_2 \ldots dx_s,$$

vom avea

$$f(y)\,dy = P\left(y < \sum_{i=1}^{s} x_i^2 \leq y + dy\right) =$$

$$= \frac{1}{(2\pi)^{\frac{s}{2}}\sigma^s} \int\int \ldots \int_{(V)} e^{-\frac{\sum_{i=1}^{s} x_i^2}{2\sigma^2}} dx_1\, dx_2 \ldots dx_s. \qquad \text{(XI.3.14)}$$

Aplicînd formula mediei, expresia (XI.3,14) devine

$$f(y)\,dy = \frac{1}{(2\pi)^{\frac{s}{2}}\sigma^s} A \cdot V, \qquad \text{(XI.3,15)}$$

unde V este domeniul de integrare egal cu volumul cuprins între cele două hipersfere cu centrele în origine şi cu razele respectiv egale cu $y^{\frac{1}{2}}$ şi $(y+dy)^{\frac{1}{2}}$, iar A, valoarea funcţiei $e^{-\frac{\sum x_i^2}{2\sigma^2}}$, într-un punct determinat din V.

Se ştie că volumul unei hipersfere, într-un spaţiu euclidian cu s dimensiuni, este dat de formula

$$\frac{\pi^{\frac{s}{2}}}{\Gamma\left(\frac{s}{2}+1\right)} \cdot R^s,$$

R fiind raza hipersferei.

Atunci, volumul cuprins între cele două hipersfere cu centrele în origine și cu razele respectiv egale cu $y^{\frac{1}{2}}$ și $(y+dy)^{\frac{1}{2}}$ va fi

$$V=\frac{\pi^{\frac{s}{2}}}{\Gamma\left(\frac{s}{2}+1\right)}\left[(y+dy)^{\frac{s}{2}}-y^{\frac{s}{2}}\right]=\frac{\pi^{\frac{s}{2}}\,s}{2\Gamma\left(\frac{s}{2}+1\right)}y^{\frac{s}{2}-1}\,dy=$$

$$=\frac{\pi^{\frac{s}{2}}}{\Gamma\left(\frac{s}{2}\right)}y^{\frac{s}{2}-1}\,dy, \qquad \text{(XI.3,16)}$$

dacă neglijăm termenii dezvoltării binomului $(y+dy)^{\frac{s}{2}}$, care conțin infiniți mici de ordin superior lui dy și înlocuim $2\Gamma\left(\frac{s}{2}+1\right)$ cu $s\Gamma\left(\frac{s}{2}\right)$.

Inlocuind în (XI.3,15) pe V cu expresia lui din (XI.3,16) și luînd $A=e^{-\frac{y}{2\sigma^2}}$, deoarece $y<\sum_{i=1}^{s}x_i^2\leq y+dy$, obținem

$$f(y)\,dy=\frac{1}{(2\pi)^{\frac{s}{2}}\cdot\sigma^s}\cdot\frac{\pi^{\frac{s}{2}}}{\Gamma\left(\frac{s}{2}\right)}\cdot y^{\frac{s}{2}-1}\cdot e^{-\frac{y}{2\sigma^2}}\cdot dy=$$

$$=\frac{1}{2^{\frac{s}{2}}\cdot\sigma^s\cdot\Gamma\left(\frac{s}{2}\right)}\cdot y^{\frac{s}{2}-1}e^{-\frac{y}{2\sigma^2}}\,dy,$$

de unde rezultă, prin schimbarea de variabilă $y=x$, expresia (XI.3,8), a funcției de frecvență a distribuției χ^2 a lui Helmert-Pearson

$$f(x)=\frac{1}{2^{\frac{s}{2}}\sigma^s\Gamma\left(\frac{s}{2}\right)}x^{\frac{s}{2}-1}\cdot e^{-\frac{x}{2\sigma^2}}\cdot$$

Distribuţia cu această funcţie de frecvenţă o vom nota pe viitor cu $H(s, \sigma)$. Ea este o distribuţie cu *s grade de libertate şi cu parametrul* σ, bucurîndu-se de proprietatea de stabilitate, aşa cum reiese din teorema de mai jos.

Teoremă. *Dacă două variabile aleatorii independente* X_1 *şi* X_2 *urmează fiecare respectiv legile* $H(s_1, \sigma)$ *cu* s_1 *grade de libertate şi* $H(s_2, \sigma)$ *cu* s_2 *grade de libertate, atunci, suma lor* X_1+X_2 *urmează legea* $H(s_1+s_2, \sigma)$ *cu* s_1+s_2 *grade de libertate.*

Demonstraţia rezultă imediat, considerînd funcţiile caracteristice ale celor două variabile X_1 şi X_2

$$\varphi_1(t)=(1-2\sigma^2 ti)^{-\frac{s_1}{2}},$$

$$\varphi_2(t)=(1-2\sigma^2 ti)^{-\frac{s_2}{2}}.$$

Funcţia caracteristică $\varphi(t)$ a variabilei X_1+X_2, în baza teoreme funcţiei caracteristice a sumei de variabile independente, va fi

$$\varphi(t)=\varphi_1(t)\cdot\varphi_2(t)=(1-2\sigma^2 ti)^{-\frac{s_1+s_2}{2}}.$$

Aceasta este expresia funcţiei caracteristice a legii $H(s_1+s_2, \sigma)$ cu s_1+s_2 grade de libertate şi teorema este astfel demonstrată.

§ 4. O proprietate de limită a distribuţiei multinomiale

La distribuţia lui Helmert-Pearson se poate ajunge, plecînd de la consideraţii diferite de cele expuse mai sus şi anume de la distribuţia multinomială.

Am văzut din studiul distribuţiei multinomiale (cap. VIII, § 3), că valoarea asimptotică a probabilităţii $P_n(\alpha_1, \alpha_2, \ldots, \alpha_s)$ este dată de

$$P_n(\alpha_1, \alpha_2, \ldots, \alpha_s)\simeq \frac{e^{-\frac{1}{2}\sum_{i=1}^{s} q_i\xi_i^2}}{(2\pi n)^{\frac{s-1}{2}}\cdot\sqrt{p_1\cdot p_2\ldots p_s}},$$

unde valorile $\alpha_1, \alpha_2, \ldots, \alpha_s$ intervin numai prin intermediul funcţiei

$$\chi^2=\sum_{i=1}^{s} q_i\xi_i^2. \qquad \text{(XI.4,1)}$$

Relația $\sum_{i=1}^{s} \alpha_i = n$ echivalează cu relația

$$\sum_{i=1}^{s} \sqrt{p_i q_i} \cdot \xi_i = 0. \qquad \text{(XI.4,2)}$$

Unei variații $\Delta\alpha_i$ a lui α_i îi corespunde lui ξ_i variația

$$\Delta\xi_i = \frac{\Delta\alpha_i}{\sqrt{np_iq_i}} \cdot$$

Deoarece relația

$$\sum_{i=1}^{s} \alpha_i = n$$

definește pe α_s în funcție de $\alpha_1, \alpha_2, \ldots, \alpha_{s-1}$, rezultă că

$$P_n(\alpha_1, \alpha_2, \ldots, \alpha_s)\,\Delta\alpha_1\Delta\alpha_2\ldots\Delta\alpha_{s-1} \simeq$$

$$\simeq \frac{\sqrt{q_1 \cdot q_2 \ldots q_{s-1}}}{(2\pi)^{\frac{s-1}{2}}\sqrt{p_s}} \cdot e^{-\frac{\chi^2}{2}} \cdot \Delta\xi_1 \cdot \Delta\xi_2 \ldots \Delta\xi_{s-1}.$$

In locul distribuției discontinue, definită de $P_n(\alpha_1, \alpha_2, \ldots, \alpha_s)$, să considerăm distribuția continuă definită de probabilitatea elementară

$$dP = \frac{\sqrt{q_1 \cdot q_2 \ldots q_{s-1}}}{(2\pi)^{\frac{s-1}{2}}\sqrt{p_s}} \cdot e^{-\frac{\chi^2}{2}} d\xi_1 \cdot d\xi_2 \ldots d\xi_{s-1}$$

împreună cu relațiile (XI.4,1) și (XI.4,2).

Ne propunem să determinăm distribuția variabilei pozitive χ^2.

In acest scop, vom folosi o demonstrație dată de T. Lișcu.

Substituția

$$\xi_i = \chi\eta_i \qquad (i = 1, 2, \ldots, s)$$

înlocuieşte variabilele ξ_i cu variabilele $\eta_1, \eta_2, \ldots, \eta_s, \chi$ legate prin relaţiile

$$\sum_{i=1}^{s} q_i \eta_i^2 = 1, \tag{XI.4,3}$$

$$\sum_{i=1}^{s} \sqrt{p_i q_i} \cdot \eta_i = 0, \tag{XI.4,4}$$

deduse din (XI.4,1) şi (XI.4,2).
Considerînd $\eta_1, \eta_2, \ldots, \eta_{s-2}, \chi$ ca variabile independente, iar η_{s-1} şi η_s ca definite din (XI.4,3) şi (XI.4,4), rezultă că

$$q_h \eta_h + q_{s-1} \eta_{s-1} \frac{\partial \eta_{s-1}}{\partial \eta_h} + q_s \eta_s \frac{\partial \eta_s}{\partial \eta_h} = 0 \qquad (h = 1, 2, \ldots, s-2), \tag{XI.4,5}$$

$$\sqrt{p_h q_h} + \sqrt{p_{s-1} q_{s-1}} \cdot \frac{\partial \eta_{s-1}}{\partial \eta_h} + \sqrt{p_s q_s} \cdot \frac{\partial \eta_s}{\partial \eta_h} = 0. \tag{XI.4,6}$$

Introducînd notaţia

$$D_i = \begin{vmatrix} q_i \eta_i & q_s \eta_s \\ \sqrt{p_i q_i} & \sqrt{p_s q_s} \end{vmatrix},$$

din (XI.4,5) şi (XI.4,6) se obţine

$$\frac{\partial \eta_{s-1}}{\partial \eta_h} = -\frac{D_h}{D_{s-1}}.$$

Din determinantul funcţional

$$\frac{D(\xi_1, \xi_2, \ldots, \xi_{s-1})}{D(\eta_1, \eta_2, \ldots, \eta_{s-2}, \chi)} = \begin{vmatrix} \chi & \ldots & 0 & \eta_1 \\ \vdots & & & \vdots \\ 0 & \ldots & \chi & \eta_{s-2} \\ \chi \frac{\partial \eta_{s-1}}{\partial \eta_1} & \ldots & \chi \frac{\partial \eta_{s-1}}{\partial \eta_{s-2}} & \eta_{s-1} \end{vmatrix} =$$

$$= \chi^{s-2} \begin{vmatrix} 1 & \ldots & 0 & \eta_1 \\ \vdots & & & \vdots \\ 0 & \ldots & 1 & \eta_{s-2} \\ \frac{\partial \eta_{s-1}}{\partial \eta_1} & \ldots & \frac{\partial \eta_{s-1}}{\partial \eta_{s-2}} & \eta_{s-1} \end{vmatrix}$$

scăzînd din ultima coloană celelalte două coloane înmulțite cu $\eta_1, \ldots, \eta_{s-2}$ respectiv, deducem

$$\frac{D(\xi_1, \xi_2, \ldots, \xi_{s-1})}{D(\eta_1, \eta_2, \ldots, \eta_{s-2}, \chi)} = \chi^{s-2}\left(\eta_{s-1} - \sum_{h=1}^{s-2} \eta_h \frac{\partial \eta_{s-1}}{\partial \eta_h}\right) = \chi^{s-2} \frac{\sum_{i=1}^{s-1} \eta_i D_i}{D_{s-1}} =$$

$$= \chi^{s-2} \frac{\begin{vmatrix} \sum_{i=1}^{s-1} q_i \eta_i^2 & q_s \eta_s \\ \sum_{i=1}^{s-1} \sqrt{q_i p_i}\, \eta_i & \sqrt{p_s q_s} \end{vmatrix}}{D_{s-1}} = \chi^{s-2} \frac{\begin{vmatrix} 1 - q_s \eta_s^2 & q_s \eta_s \\ -\sqrt{p_s q_s}\, \eta_s & \sqrt{p_s q_s} \end{vmatrix}}{D_{s-1}} = \chi^{s-2} \frac{\sqrt{p_s q_s}}{D_{s-1}}.$$

Prin substituția

$$\xi_i = \chi \eta_i \qquad (i = 1, 2, \ldots, s)$$

fiecărui punct $(\xi_1, \xi_2, \ldots, \xi_{s-1})$ îi corespunde un singur punct $(\eta_1, \eta_2, \ldots, \eta_{s-2}, \chi)$ $(\chi > 0)$, obținut prin determinarea lui ξ_s din (XI.4,2), a lui χ din (XI.4,1) și a lui η_i din această substituție.

Reciproc, fiecărui punct $(\eta_1, \eta_2, \ldots, \eta_{s-2}, \chi)$, $(\chi > 0)$ cu

$$\sum_{h=1}^{s-2} q_h \eta_h^2 < 1,$$

îi corespund două puncte $(\xi_1, \xi_2, \ldots, \xi_{s-1})$, obținute prin determinarea a două sisteme de valori (η_{s-1}, η_s) din (XI.4,1) (XI.4,4) și substituția $\xi_i = \chi \eta_i$.

Deci, prin substituția $\xi_i = \chi \eta_i$, spațiul $\xi_1, \xi_2, \ldots, \xi_{s-1}$ acoperă de două ori spațiul $\eta_1, \eta_2, \ldots, \eta_{s-1}$.

In fiecare dintre cele două puncte $(\xi_1, \xi_2, \ldots, \xi_{s-1})$ care corespund aceluiași punct $(\eta_1, \eta_2, \ldots, \eta_{s-2}, \chi)$,

$$\left| \frac{D(\xi_1, \xi_2, \ldots, \xi_{s-1})}{D(\eta_1, \eta_2, \ldots, \eta_{s-2}, \chi)} \right| = \chi^{s-2} \cdot \frac{\sqrt{p_s q_s}}{|D_{s-1}|}$$

are aceeași valoare, deoarece

$$D_{s-1}^2 = q_{s-1} q_s \left[(p_{s-1} + p_s)(q_{s-1} \eta_{s-1}^2 + q_s \eta_s^2) - (\sqrt{p_{s-1} q_{s-1}} \cdot \eta_{s-1} + \right.$$

$$\left. + \sqrt{p_s q_s}\, \eta_s)^2\right] = q_{s-1} q_s \left[(p_{s-1} + p_s)\left(1 - \sum_{h=1}^{s-2} q_h \eta_h^2\right) - \left(\sum_{h=1}^{s-2} \sqrt{p_h q_h}\, \eta_h^2\right)\right]$$

şi în consecinţă expresiile $|D_{s-1}|$ sînt unic determinate în funcţie de $\eta_1, \eta_2, \ldots, \eta_{s-2}$.
Rezultă că

$$dP = 2\frac{\sqrt{q_1\, q_2 \ldots q_{s-1}}}{(2\pi)^{\frac{s-1}{2}}\sqrt{p_s}} e^{-\frac{\chi^2}{2}} \cdot \chi^{s-2} \cdot \frac{\sqrt{p_s q_s}}{|D_{s-1}|} d\eta_1 \cdot d\eta_2 \ldots d\eta_{s-2} \cdot d\chi.$$

Deoarece

$$\chi^{s-2} \cdot e^{\frac{\chi^2}{2}} > 0,$$

pentru orice $\chi > 0$,

$$F(x) = \frac{\int_0^x \chi^{s-2} e^{-\frac{\chi^2}{2}} d\chi}{\int_0^\infty \chi^{s-2} e^{-\frac{\chi^2}{2}} d\chi}$$

este o funcţie de distribuţie.
Prin schimbarea de variabile $t = \frac{\chi^2}{2}$, găsim

$$\int_0^\infty \chi^{s-2} e^{-\frac{\chi^2}{2}} d\chi = 2^{\frac{s-3}{2}} \int_0^\infty t^{\frac{s-3}{2}} e^{-t} dt = 2^{\frac{s-3}{2}} \Gamma(s-1).$$

Aşadar

$$dP = \frac{\chi^{s-2} \cdot e^{-\frac{\chi^2}{2}} \cdot d\chi}{2^{\frac{s-3}{2}} \cdot \Gamma\left(\frac{s-1}{2}\right)} \cdot \frac{\Gamma\left(\frac{s-1}{2}\right)}{\pi^{\frac{s-1}{2}}} \cdot \frac{\sqrt{q_1 \cdot q_2 \ldots q_s}}{|D_{s-1}|} d\eta_1 \cdot d\eta_2 \ldots d\eta_{s-2}$$

este produsul a două probabilităţi elementare independente, dintre care prima este legea lui χ^2.
Introducînd notaţia

$$\sum_{i=1}^{s} \frac{(\alpha_i - np_i)^2}{np_i} = \chi^2,$$

rezultă că probabilitatea elementară de a avea pentru expresia de mai sus o probabilitate cuprinsă între χ şi $\chi+d\chi$ este egală cu

$$\frac{\chi^{s-2}\cdot e^{-\frac{\chi^2}{2}}\cdot d\chi}{2^{\frac{s-3}{2}}\cdot\Gamma\left(\frac{s-1}{2}\right)}. \qquad \text{(XI.4,7)}$$

Se observă uşor că expresia (XI.4,7) este funcţia de frecvenţă $f(x)\,dx$ a distribuţiei $H(s-1, 1)$ a lui Helmert-Pearson cu $s-1$ grade de libertate ($\sigma=1$).

In adevăr, făcînd substituţia $\chi^2=x$, expresia (XI.4,7) de mai sus devine

$$f(x)\,dx=\frac{1}{2^{\frac{s-1}{2}}\cdot\Gamma\left(\frac{s-1}{2}\right)}\cdot x^{\frac{s-1}{2}-1}\cdot e^{-\frac{x}{2}}\cdot dx.$$

§ 5. Distribuţii date de forme pătratice de variabile aleatorii

Să considerăm forma pătratică

$$Q(x)=\sum_{k=1}^{n} x_k^2-\sum_{k=1}^{n} l_k^2(x) \qquad (s<n), \qquad \text{(IX.5,1)}$$

unde x_k sînt variabile independente, care urmează legea normală $N(0, 1)$ iar $l_k(x)$ sînt forme liniare în x_i, independente, care, de asemenea, urmează legea normală $N(0, 1)$:

$$l_k(x)=\sum_{i=1}^{n} a_{ki}\,x_i.$$

In acest caz, vom arăta că variabila $Q(x)$ dată de expresia (XI.5,1) dă naştere la distribuţia $H(n-s, 1)$ a lui Helmert-Pearson cu $n-s$ grade de libertate şi parametrul $\sigma=1$.

Deoarece $l_j(x), l_k(x)$ sînt independente, rezultă că

$$M[l_j(x)\cdot l_k(x)]=M(l_j)\cdot M(l_k)=\sum_{i=1}^{n} a_{j_i}a_{k_i}=0 \qquad (j,\ k=1,\ 2,\ldots,\ s),$$

$$M[l_j^2(x)]=\sum_{i=1}^{n} a_{k_i}^2=1.$$

Să introducem alte $n-s$ forme liniare

$$l_{s+1},\ l_{s+2},\ldots,\ l_n,$$

unde

$$l_k(x)=\sum_{i=1}^{n} a_{k_i}x_i \qquad (s<k\leqslant n)$$

şi coeficienţii a_{k_i} satisfăcînd relaţiile de ortogonalitate

$$\sum_{i=1}^{n} a_{k_i}=0,\ \sum_{i=1}^{n} a_{j_i}a_{k_i}=0 \qquad (j,\ k=1,\ 2,\ldots,\ n).$$

Variabilele

$$l_1,\ l_2,\ldots,\ l_s,\ldots,\ l_n$$

fiind sume de variabile normale sînt normale şi, fiind necorelate din cauza normalităţii, sînt şi independente.
Să aplicăm transformarea liniară

$$y_k=\sum_{i=1}^{n} a_{k_i}x_i. \qquad \text{(XI.5,2)}$$

Atunci, expresia (XI.5,1) a lui $Q(x)$ devine

$$Q(x)=\sum_{k=1}^{n} x_k^2-\sum_{k=1}^{s} l_k^2(x)=\sum_{k=1}^{n}(x_k^2-y_k^2)+$$

$$+\sum_{k=s+1}^{n} y_k^2=\sum_{k=s+1}^{n} y_k^2=\sum_{k=s+1}^{n} l_k^2(x), \qquad \text{(XI.5,3)}$$

întrucît transformarea (XI.5,2) este ortogonală şi $y_k=l_k(x)$. Variabilele $l_k(x)$ urmînd fiecare legea normală $N(0,\ 1)$ şi fiind, în acelaşi timp, independente, înseamnă că variabila (XI.5,3)

$$Q(x)=\sum_{k=s+1}^{n} l_k^2(x)$$

dă naștere distribuției $H(n-s, 1)$ cu $n-s$ grade de libertate și parametrul $\sigma=1$.

Se înțelege că această proprietate se păstrează și pentru variabilele x_k și $l_k(x)$ care urmează legea normală $N(0, \sigma)$.

De observat că variabila $Q(x)=\sum_{k=s+1}^{n} l_k^2(x)$ este independentă de variabilele

$$l_1(x),\ l_2(x),\ldots,\ l_s(x)$$

și că rangul formei pătratice $Q(x)$ coincide cu numărul gradelor de libertate.

Variabila dă naștere, de altminteri, la o interpretare geometrică simplă. Astfel, punctul $P(x_1, x_2, \ldots, x_n)$ fiind un punct într-un spațiu euclidian R^n, expresia $\sum_{k=1}^{n} x_k^2$ reprezintă suma pătratelor distanțelor de la punctul P la fețele n-edrului ortogonal, care a fost ales ca sistem de referință, iar $\sum_{k=1}^{s} l_k^2(x)$, suma pătratelor distanțelor de la punctul P la S din fețele unui n-edru ortogonal cu vîrful în origine.

Deoarece se știe că suma pătratelor distanțelor de la un punct la cele n fețe ale unui n-edru ortogonal este constantă, avem

$$\sum_{k=1}^{n} x_k^2-\sum_{k=1}^{s} l_k^2(x)=\sum_{k=s+1}^{n} l_k^2(x).$$

De aici, se vede pentru care motiv forma considerată este de rangul $n-s$.

Teoremă. *Fie n variabile independente x_i, fiecare urmînd legea normală $N(0, \sigma)$ și $Q_1, Q_2, \ldots, Q_s$ forme pătratice în x_i respectiv de rangul $k_1, k_2, \ldots, k_s$.*

Dacă avem

$$\sum_{h=1}^{s} Q_h=\sum_{h=1}^{n} x_h^2,$$

atunci condiția necesară și suficientă ca distribuțiile Q_h $(h=1, 2, \ldots s)$ să fie independente este

$$k_1+k_2+\ldots+k_s=n.$$

In acest caz, distribuţia variabilei Q_h *este* $H(k_h, \sigma)$.
Demonstraţia acestei teoreme, datorită lui Cochran, este imediată. Condiţia este necesară, pentru că, dacă variabilele Q_h sînt independente şi urmează, respectiv, legea $H(k_h, \sigma)$, atunci $\sum_{h=1}^{s} Q_h$ urmează legea $H\left(\sum_{h=1}^{s} k_h, \sigma\right)$, iar $\sum_{h=1}^{n} x_h^2$ legea $H(n, \sigma)$.
Trebuie, deci, să avem

$$\sum_{k=1}^{s} k_h = n.$$

Condiţia este şi suficientă, deoarece, descompunerea în sumă algebrică de pătrate a formei pătratice ne dă

$$\sum_{h=1}^{s} Q_h = \sum_{h=1}^{n} x_h^2 = \sum_{h=1}^{n} l_h^2(x).$$

Transformarea

$$y_h = l_h(x) \qquad (h = 1, 2, \ldots, n),$$

lăsînd invariantă forma $\sum_{h=1}^{n} x_h^2$, este ortogonală, de unde rezultă imediat că $l_h(x)$ sînt independente şi urmează legea normală $N(0, \sigma)$. Cum, însă, Q_h este o sumă de k_h pătrate $l_h(x)$, deducem că variabilele Q_h sînt independente şi urmează legea $H(k_h, \sigma)$.

§ 6. Reprezentarea grafică a distribuţiei $H(s, \sigma)$

Reprezentarea grafică a funcţiei de frecvenţă

$$f(x) = \frac{1}{2^{\frac{s}{2}} \sigma^s \Gamma\left(\frac{s}{2}\right)} \cdot x^{\frac{s}{2}-1} \cdot e^{-\frac{x^2}{2\sigma^2}}$$

a distribuţiei $H(s, \sigma)$ cu s grade de libertate şi parametrul σ este sugestivă.

Scanned with OKEN Scanner

Domeniul de determinare al acestei funcții este mulțimea valorilor lui x, care satisfac inegalitățile

$$0 \leqq x < +\infty,$$

deoarece variabila x este o sumă de s pătrate ($x = x_1^2 + x_2^2 + \ldots + x_s^2$) și, prin urmare, funcția $f(x)$ ia numai valori pozitive.

Este interesant de a studia reprezentarea grafică a distribuției $H(s, \sigma)$ pentru următoarele valori ale lui s.

1) Dacă $s=1$, avem distribuția $H(1, \sigma)$ a cărei funcție de frecvență este egală cu

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} x^{-\frac{1}{2}} \cdot e^{-\frac{x}{2\sigma^2}}. \qquad \text{(XI.6,1)}$$

Se vede imediat că, pentru $x=0$, $f(x) \to \infty$, iar cînd $x \to \infty$, $f(x)=0$. In intervalul de existență, funcția (XI.6,1) este descrescătoare și are concavitatea îndreptată în sus.

In adevăr, derivînd, găsim

$$f'(x) = -\frac{1}{2\sqrt{2\pi}\cdot\sigma} \cdot x^{-\frac{1}{2}} \cdot e^{-\frac{x}{2\sigma^2}} (x^{-1} + \sigma^{-2}) =$$

$$= -\frac{1}{2} f(x) (x^{-1} + \sigma^{-2}) < 0$$

și

$$f''(x) = -\frac{1}{2} f'(x) (x^{-1} + \sigma^{-2}) + \frac{1}{2} f(x) (x^{-1} + \sigma^{-2}) > 0,$$

pentru orice $0 \leqq x < \infty$.

Funcția (XI.6,1) nu are nici maxim, nici minim.

2) Dacă $s=2$, obținem distribuția $H(2, \sigma)$, care are funcția de frecvență

$$f(x) = \frac{1}{2\sigma^2} \cdot e^{-\frac{x}{2\sigma^2}}. \qquad \text{(XI.6,2)}$$

Observăm că, pentru $x=0$, $f(x) = \frac{1}{2\sigma^2}$ și cînd $x \to \infty$, $f(x)=0$. De asemenea, în intervalul de existență al ei, funcția (XI.6,2) este des-

crescătoare şi are concavitatea în direcţia y-lor pozitivi, ceea ce se constată din semnul derivatei întîia şi a doua.
Astfel

$$f'(x)=-\frac{1}{4\sigma^4}e^{-\frac{x}{2\sigma^2}}=-\frac{1}{2\sigma^2}f(x)<0 \qquad (0\leq x<\infty)$$

$$f''(x)=-\frac{1}{2\sigma^2}f'(x)>0.$$

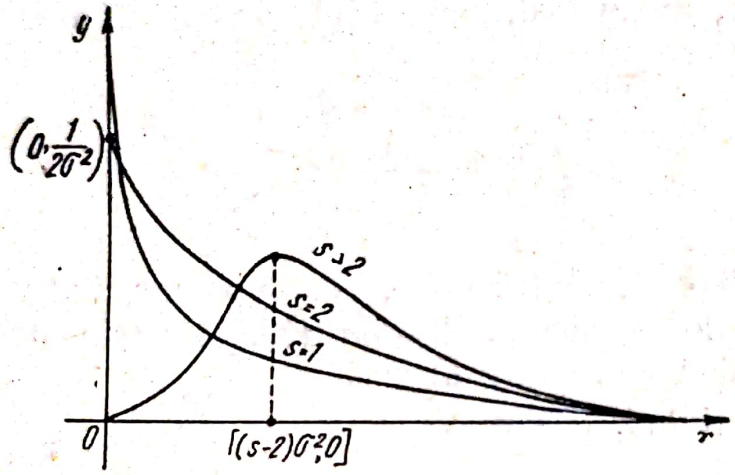


Fig. XI.6,1

Funcţia $\frac{1}{2\sigma^2}e^{-\frac{x}{2\sigma^2}}$ are, deci, un maxim egal cu $\frac{1}{2\sigma^2}$.

3) In sfîrşit, dacă $s>2$, vom avea de examinat distribuţia $H(s,\sigma)$ cu funcţia de frecvenţă

$$f(x)=\frac{1}{2^{\frac{s}{2}}\cdot\sigma^s\cdot\Gamma\left(\frac{s}{2}\right)}\cdot x^{\frac{s}{2}-1}\cdot e^{-\frac{x}{2\sigma^2}} \qquad (s>2).$$

Pentru $x=0$, $f(x)=0$, iar cînd $x\to\infty$, $f(x)\to 0$.
Prin derivare obţinem

$$f'(x)=\frac{1}{2^{\frac{s}{2}}\sigma^s\Gamma\left(\frac{s}{2}\right)}x^{\frac{s}{2}-1}\cdot e^{-\frac{x}{2\sigma^2}}\cdot\left[\left(\frac{s}{2}-1\right)-\frac{1}{2\sigma^2}\right]=$$

$$=f(x)\left(\frac{s-2}{2}x^{-1}-2\sigma^{-2}\right).$$

Derivata se anulează în punctul de abscisă $x=(s-2)\sigma^2$ și își schimbă semnul în acest punct, trecînd de la valori pozitive la valori negative.

Deci, funcția $f(x)$ are un maxim pentru $x=(s-2)\sigma^2$.

Graficul funcției de frecvență al distribuției $H(s, \sigma)$ pentru cele trei cazuri se arată în fig. XI.6,1.

§ 7. Momentele distribuției $H(s, \sigma)$

Momentele distribuției $H(s, \sigma)$ rezultă imediat din derivatele funcției sale caracteristice (XI.2,11) în raport cu t.

Astfel

$$\varphi(t)=(1-2\sigma^2 ti)^{-\frac{s}{2}},$$

$$\varphi'(t)=is\sigma^2(1-2\sigma^2 ti)^{-\frac{s}{2}-1},$$

$$\varphi''(t)=i^2 s(s+2)\sigma^4(1-2\sigma^2 ti)^{-\frac{s}{2}-2},$$

. . . . . .

$$\varphi^{(r)}(t)=i^r s(s+2)\ldots(s+2r-2)\sigma^{2r}\cdot(1-2\sigma^2 ti)^{-\frac{s}{2}-r},$$

de unde făcînd $t=0$, deducem

$$M_r=\frac{1}{i^r}\cdot\left[\frac{d^r\varphi(t)}{dt^r}\right]_{t=0}=s(s+2)\ldots[s+2(r-1)]\cdot\sigma^{2r}, \quad (XI.7,1)$$

adică

$$M_1=s\sigma^2,$$
$$M_2=s(s+2)\sigma^4,$$
$$M_3=s(s+2)(s+4)\sigma^6,$$
$$M_4=s(s+2)(s+4)(s+6)\sigma^8,$$
. . . . . .

Scanned with OKEN Scanner

$$m_1 = 0,$$

$$m_2 = M_2 - M_1^2 = 2s\sigma^4,$$

$$m_3 = M_3 - 3M_2M_1 + 2M_1^3 = 8s\sigma^6,$$

$$m_4 = M_4 - 4M_3M_1 + 6M_2M_1^2 - 3M_1^4 = 12s\sigma^8(s+4),$$

. . . . . .

Cunoscînd momentele M_1 şi m_2 se poate arăta acum că distribuţia $H(s, \sigma)$ se apropie asimptotic de distribuţia normală, cînd numărul gradelor de libertate s este foarte mare.

Teoremă. *Dacă* $s \to \infty$, *atunci distribuţia* $H(s, \sigma)$ *cu* s *grade de libertate tinde către distribuţia normală* $N(0, 1)$.

Pentru demonstrarea teoremei ne vom folosi de funcţia caracteristică a variabilei reduse

$$\frac{x - M_1(x)}{\sqrt{m_2}} = \frac{x - s\sigma^2}{\sqrt{2s\,\sigma^2}},$$

care, în baza definiţiei sale, se va scrie

$$\Psi_s(t) = M\left(e^{\frac{x - s\sigma^2}{\sqrt{2s}\cdot\sigma^2} ti}\right) = e^{-\sqrt{\frac{s}{2}}\, ti} \cdot M\left[e^{x\left(\frac{t}{\sqrt{2s}\,\sigma^2}\right) t}\right] =$$

$$= e^{-\sqrt{\frac{s}{2}}\, ti} \cdot \varphi\left(\frac{t}{\sqrt{2s}\cdot\sigma^2}\right) = e^{-\sqrt{\frac{s}{2}}\, ti}\left(1 - \sqrt{\frac{s}{2}} \cdot ti\right)^{-\frac{s}{2}}. \qquad \text{(XI. 7,2)}$$

Aplicînd logaritmii în expresia (XI. 7,2), găsim

$$\ln \Psi_s(t) = -\sqrt{\frac{s}{2}} \cdot ti - \frac{s}{2} \ln\left(1 - \sqrt{\frac{2}{s}} \cdot ti\right). \qquad \text{(XI. 7,3)}$$

Observăm că logaritmul din (XI. 7,3) se poate dezvolta în serie, dacă

$$\left|\sqrt{\frac{2}{s}} \cdot t\right| < 1,$$

Deci,

$$\ln \Psi_s(t) = -\sqrt{\frac{s}{2}} \cdot ti + \frac{s}{2}\left[\sqrt{\frac{2}{s}}\, ti - \frac{t^2}{2} + O\left(s^{-\frac{3}{2}}\right)\right] = \left[-\frac{t^2}{2} + O\left(s^{-\frac{1}{2}}\right)\right],$$

$O\left(s^{-\frac{1}{2}}\right)$ fiind o funcţie asimptotică de ordinul lui $s^{-\frac{1}{2}}$.

Scanned with OKEN Scanner

La limită, cînd $s\to\infty$, $O\left(s^{-\frac{1}{2}}\right)=0$ şi, prin urmare, $\Psi_s(t)\to e^{-\frac{t^2}{2}}$, care este funcţia caracteristică a legii normale $N(0,1)$.
Se ştie că, dacă şirul de funcţii caracteristice $\Psi_s(t)$ tinde către funcţia caracteristică $e^{-\frac{t^2}{2}}$ a legii normale $N(0,1)$, atunci şi funcţiile de repartiţie corespunzătoare vor tinde, de asemenea, către funcţia de repartiţie a legii normale $N(0,1)$.
Teorema este, astfel, demonstrată.

§ 8. Distribuţii înrudite cu distribuţia $H(s,\sigma)$

Am arătat că distribuţia $H(s,\sigma)$ a lui Helmert-Pearson cu s grade de libertate şi cu parametrul σ este dată de variabila $\sum_{i=1}^{s} x_i^2$, unde fiecare dintre variabilele x_i urmează legea normală $N(0,\sigma)$.
Ne propunem să determinăm funcţiile de frecvenţă ale distribuţiilor medii, date de variabilele

$$\frac{\sum_{i=1}^{s} x_i^2}{s},\quad \sqrt{\sum_{i=1}^{s} x_i^2},\quad \sqrt{\frac{\sum_{i=1}^{s} x_i^2}{s}},$$

care se întîlnesc frecvent în Statistica matematică.
In acest scop, determinăm funcţiile de repartiţie ale acestor variabile.

1) Variabila $\frac{\sum_{i=1}^{s} x_i^2}{s}$ are ca funcţie de repartiţie

$$F_1(x)=P\left(\frac{\sum_{i=1}^{s} x_i^2}{s}\leqslant x\right)=P\left(\sum_{i=1}^{s} x_i^2\leqslant sx\right)=H(sx),$$

dacă notăm $P\left(\sum_{i=1}^{s} x_i^2\leqslant x\right)=H(x)$.

Funcţia de frecvenţă corespunzătoare este

$$f_1(x) = s \cdot h(sx), \tag{XI. 8,1}$$

unde $h(x) = \frac{dH(x)}{dx}$.

Observăm că $h(x)$ este funcţia de frecvenţă a distribuţiei $H(s, \sigma)$, deci, după cum am văzut mai înainte, are expresia

$$h(x) = \frac{1}{2^{\frac{s}{2}} \cdot \sigma^s \cdot \Gamma\left(\frac{s}{2}\right)} \cdot x^{\frac{s}{2}-1} \cdot e^{-\frac{x}{2\sigma^2}}.$$

Atunci, expresia (XI. 8,1) de mai sus devine

$$f_1(x) = \frac{s^{\frac{s}{2}}}{2^{\frac{s}{2}} \cdot \sigma^s \Gamma\left(\frac{s}{2}\right)} \cdot x^{\frac{s}{2}-1} \cdot e^{-\frac{xs}{2\sigma^2}}; \tag{XI. 8.2}$$

$f_1(x)$ este funcţia de frecvenţă a distribuţiei variabilei $\frac{\sum_{i=1}^{s} x_i^2}{s}$.

2) Calculăm acum funcţia de repartiţie a variabilei $\sqrt{\sum_{i=1}^{s} x_i^2}$

$$F_2(x) = P\left(\sqrt{\sum_{i=1}^{s} x_i^2} \leqslant x\right) = P\left(\sum_{i=1}^{s} x_i^2 \leqslant x^2\right) = H(x^2),$$

de unde, derivînd în raport cu x, deducem

$$f_2(x) = 2xh(x^2) = \frac{2}{2^{\frac{s}{2}} \cdot \sigma^s \Gamma\left(\frac{s}{2}\right)} \cdot x^{s-1} e^{-\frac{x^2}{2\sigma^2}}. \tag{XI.8,3}$$

Această distribuţie mai este cunoscută sub numele de *distribuţia lui Maxwell*, care a pus-o în valoare prin aplicaţiile ei în fizică.

Aşa, de exemplu, dacă $x_i\ (i = 1, 2, \ldots, s)$ sînt proiecţiile vitezei v ale unei molecule dintr-un gaz, în spaţiul cu s dimensiuni, atunci

$$v = \sqrt{x_1^2 + x_2^2 + \ldots + x_s^2}.$$

Scanned with OKEN Scanner

In ipoteza că fiecare proiecție x_i urmează o lege normală $N(0, \sigma)$, rezultă că viteza moleculei din acel gaz urmează o lege (XI.8,3)

$$\frac{2}{2^{\frac{s}{2}} \cdot \sigma^s \cdot \Gamma\left(\frac{s}{2}\right)} \cdot x^{s-1} \cdot e^{-\frac{x^2}{2\sigma^2}}.$$

3) Ne-a mai rămas să determinăm distribuția variabilei $\sqrt{\frac{\sum_{i=1}^{s} x_i^2}{s}}$, a cărei funcție de repartiție se va scrie

$$F_3(x) = P\left(\sqrt{\frac{\sum_{i=1}^{s} x_i^2}{s}} \leqslant x\right) = P\left(\sum_{i=1}^{s} x_i^2 \leqslant sx^2\right) = H(sx^2).$$

In sfirșit, derivînd pe $F_3(x)$ în raport cu x, vom găsi funcția de frecvență a distribuției variabilei considerate

$$f_3(x) = 2sxh(sx^2) = \frac{2s^{\frac{s}{2}}}{2^{\frac{s}{2}} \cdot \sigma^s \cdot \Gamma\left(\frac{s}{2}\right)} \cdot x^{s-1} \cdot e^{-\frac{sx^2}{2\sigma^2}}.$$

§ 9. Distribuția Student

Să considerăm variabila dată de raportul

$$t = \frac{X}{\sqrt{\frac{Y}{s}}}, \qquad \text{(XI.9,1)}$$

unde X și Y sînt două variabile care urmează, respectiv, legile de probabilitate $N(0, \sigma)$ și $H(s, \sigma)$.

1) Probabilitatea elementară pentru punctul (X, Y) este

$$\frac{1}{\sqrt{\pi} \cdot 2^{\frac{s+1}{2}} \cdot \sigma^{s+1} \cdot \Gamma\left(\frac{s}{2}\right)} Y^{\frac{s}{2}-1} \cdot e^{-\frac{X^2+Y}{2\sigma^2}} \cdot dX \cdot dY. \qquad \text{(XI.9,2)}$$

Efectuînd schimbarea de variabile

$$X=t\sqrt{\frac{u}{s}} \quad \text{şi} \quad Y=u,$$

determinantul funcţional al transformării are valoarea

$$\frac{D(X,Y)}{D(t,u)}=\begin{vmatrix}\sqrt{\frac{u}{s}} & \frac{t}{2\sqrt{us}}\\ 0 & 1\end{vmatrix}=\sqrt{\frac{u}{s}}.$$

Atunci, probabilitatea elementară (XI.9,2) pentru punctul (X, Y) devine

$$\frac{1}{\sqrt{\pi s}\cdot 2^{\frac{s+1}{2}}\cdot\sigma^{s+1}\cdot\Gamma\left(\frac{s}{2}\right)}\,u^{\frac{s-1}{2}}\cdot e^{-u\frac{\frac{t^2}{s}+1}{2\sigma^2}}\cdot dt\cdot du.$$

Pentru a găsi distribuţia marginală a lui t, vom integra expresia de mai sus în raport cu variabila u între limitele 0 şi ∞.
Astfel vom avea

$$g(t)\,dt=\frac{dt}{\sqrt{\pi s}\cdot 2^{\frac{s+1}{2}}\cdot\sigma^{s+1}\cdot\Gamma\left(\frac{s}{2}\right)}\int_0^{\infty}u^{\frac{s-1}{2}}e^{-u\frac{\frac{t^2}{s}+1}{2\sigma^2}}du. \qquad \text{(XI.9,3)}$$

Ţinînd seama de definiţia funcţiei $\Gamma(x)$ deducem

$$\int_0^{\infty}u^{\frac{s-1}{2}}\cdot e^{-u\frac{1+\frac{t^2}{s}}{2\sigma^2}}\cdot du=\frac{\Gamma\left(\frac{s+1}{2}\right)}{\left(\frac{1+\frac{t^2}{s}}{2\sigma^2}\right)^{\frac{s+1}{2}}}=2^{\frac{s+1}{2}}\cdot\sigma^{s+1}\cdot\Gamma\left(\frac{s+1}{2}\right)\left(1+\frac{t^2}{s}\right)^{-\frac{s+1}{2}}.$$

Prin urmare, expresia (XI. 9,3) devine

$$g(t)=\frac{\Gamma\left(\frac{s+1}{2}\right)}{\sqrt{\pi s}\cdot\Gamma\left(\frac{s}{2}\right)}\cdot\left(1+\frac{t^2}{s}\right)^{-\frac{s+1}{2}}, \qquad \text{(XI.9,4)}$$

care este funcţia de frecvenţă a distribuţiei variabilei (XI.9,1).

Scanned with OKEN Scanner

După cum se constată, expresia de mai sus $g(t)$ este independentă de parametrul σ. Această observaţie are mare importanţă în aplicaţii. Distribuţia astfel determinată (XI.9,4) este cunoscută sub numele de *distribuţia Student cu s grade de libertate.*

2) Funcţia de frecvenţă (XI.9,4) a distribuţiei Student cu s grade de libertate se poate, de asemenea, determina scriind variabila (XI.9,1) sub forma

$$t=\frac{x_{s+1}}{\sqrt{\frac{x_1^2+x_2^2+\ldots+x_s^2}{s}}}, \qquad \text{(XI.9,5)}$$

unde fiecare dintre variabilele $x_i\,(i=1,2,\ldots,s+1)$ este o variabilă normală $N(0,\sigma)$.

Să presupunem că $x_1, x_2, \ldots, x_{s+1}$ sînt coordonatele unui punct $M(x_1, x_2, \ldots, x_{s+1})$ dintr-un spaţiu euclidian cu $s+1$ dimensiuni.

In acest caz, probabilitatea elementară pentru punctul $M(x_1, x_2, \ldots \ldots, x_{s+1})$ se va scrie

$$\frac{1}{(2\pi)^{\frac{s+1}{2}}\cdot\sigma^{s+1}}\cdot e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{s+1}x_i^2}\cdot dx_1\cdot dx_2\ldots dx_{s+1}. \qquad \text{(XI.9,6)}$$

Facem, acum, următoarea transformare de variabile în coordonate polare

$$\begin{aligned}
x_1&=\rho\cos\theta_1\cdot\cos\theta_2\ldots\cos\theta_s,\\
x_2&=\rho\cos\theta_1\cdot\cos\theta_2\ldots\cos\theta_{s-1}\cdot\sin\theta_s,\\
x_3&=\rho\cos\theta_1\cdot\cos\theta_2\ldots\cos\theta_{s-2}\cdot\sin\theta_{s-1},\\
&\ldots\ldots\ldots\ldots\ldots\ldots\\
x_s&=\rho\cos\theta_1\sin\theta_2,\\
x_{s+1}&=\rho\sin\theta_1.
\end{aligned} \qquad \text{(XI.9,7)}$$

In urma transformării (XI.9,7), expresia (XI.9,6) devine

$$\frac{1}{(2\pi)^{\frac{s+1}{2}}\cdot\sigma^{s+1}}\cdot e^{-\frac{\rho^2}{2\sigma^2}}\cdot\rho^s\,|\cos^{s-1}\theta_1\cos^{s-2}\theta_2\ldots\cos\theta_{s-1}|\,d\rho\cdot d\theta_1\cdot d\theta_2\ldots d\theta_s, \qquad \text{(XI.9,8)}$$

avînd în vedere că probabilitatea este o funcţie pozitivă.

Din transformarea de variabile (XI.9,7) deducem

$$\sqrt{x_1^2+x_2^2+\ldots+x_s^2}=\rho\cos\theta_1$$

şi deci, variabila (XI.9,5) se poate scrie

$$t=\sqrt{s}\cdot\operatorname{tg}\theta_1. \tag{XI.9,9}$$

Integrînd expresia (XI.9,5) în raport cu variabilele

$$\rho,\theta_2,\theta_3,\ldots,\theta_s,$$

determinăm distribuţia marginală a lui θ_1

$$f(\theta_1)\,d\theta_1=$$

$$=2^s\,\frac{\cos^{s-1}\theta_1\,d\theta_1}{(2\pi)^{\frac{s+1}{2}}\cdot\sigma^{s+1}}\cdot\int_0^\infty\rho^s\cdot e^{-\frac{\rho^2}{2\sigma^2}}d\rho\int_0^{\frac{\pi}{2}}\cos^{s-2}\theta_2\,d\theta_2\ldots\int_0^{\frac{\pi}{2}}\cos\theta_{s-1}\,d\theta_{s-1}\int_0^{\frac{\pi}{2}}d\theta_s. \tag{XI.9,10}$$

Factorul 2^s rezultă din faptul că integrarea s-a făcut între limitele 0 şi $\frac{\pi}{2}$ în spaţiul cu s dimensiuni.

Calcularea expresiei (XI.9,10) se face uşor, observînd pe de o parte că integrala $\int_0^\infty\rho^s\cdot e^{-\frac{\rho^2}{2\sigma^2}}d\rho$ se reduce la o funcţie $\Gamma(x)$ prin substituţia $\rho^2=u$.

Intr-adevăr, avem

$$\int_0^\infty\rho^s e^{-\frac{\rho^2}{2\sigma^2}}d\rho=\frac{1}{2}\int_0^\infty u^{\frac{s-1}{2}}e^{-\frac{u}{2\sigma^2}}du=\frac{2^{\frac{s+1}{2}}\cdot\sigma^{s+1}\cdot\Gamma\left(\frac{s+1}{2}\right)}{2}.$$

Pe de altă parte, se ştie că

$$\int_0^{\frac{\pi}{2}}\cos^m x=\frac{\sqrt{\pi}}{2}\cdot\frac{\Gamma\left(\frac{m+1}{2}\right)}{\Gamma\left(\frac{m+2}{2}\right)}.$$

Scanned with OKEN Scanner

Atunci, expresia (XI.9,10) devine

$$f(\theta_1)\,d\theta_1 = 2^s \cdot \frac{\cos^{s-1}\theta_1 d\theta_1}{(2\pi)^{\frac{s+1}{2}} \cdot \sigma^{s+1}} \cdot \frac{2^{\frac{s+1}{2}} \cdot \sigma^{s+1} \cdot \Gamma\left(\frac{s+1}{2}\right)}{2} \cdot \frac{\pi^{\frac{s-2}{2}}}{2^{s-2}} \cdot \frac{\Gamma\left(\frac{s-1}{2}\right)}{\Gamma\left(\frac{s}{2}\right)} \times$$

$$\times \frac{\Gamma\left(\frac{s-2}{2}\right)}{\Gamma\left(\frac{s-1}{2}\right)} \cdots \frac{\Gamma(1)}{\Gamma\left(\frac{3}{2}\right)} \cdot \frac{\pi}{2} = \frac{\Gamma\left(\frac{s+1}{2}\right)}{\sqrt{\pi} \cdot \Gamma\left(\frac{s}{2}\right)} \cdot \cos^{s-1}\theta_1\, d\theta_1, \quad \text{(XI.9,11)}$$

adică legea de probabilitate a lui θ_1.

Deoarece ne interesează legea de probabilitate a variabilei t, vom trece de la variabila θ_1 la variabila t, folosind relația (XI.9,9) stabilită mai sus.

Astfel, vom avea

$$t = \sqrt{s} \cdot \operatorname{tg}\theta_1$$

sau

$$dt = \sqrt{s} \cdot \frac{d\theta_1}{\cos^2\theta_1},$$

de unde rezultă

$$\cos\theta_1 = \left(1 + \frac{t^2}{s}\right)^{-\frac{1}{2}}$$

$$d\theta_1 = \frac{1}{\sqrt{s}}\left(1 + \frac{t^2}{s}\right)^{-1} \cdot dt.$$

In sfîrșit, înlocuind valorile obținute pentru $\cos\theta_1$ și $d\theta_1$ în expresia (XI.9,11), vom regăsi legea de probabilitate a variabilei t (XI.9,4)

$$g(t) = \frac{\Gamma\left(\frac{s+1}{2}\right)}{\sqrt{\pi s} \cdot \Gamma\left(\frac{s}{2}\right)}\left(1 + \frac{t^2}{s}\right)^{-\frac{s+1}{2}} \cdot$$

Scanned with OKEN Scanner

§ 10. Momentele distribuției Student

Am arătat că funcția de frecvență a distribuției Student cu s grade de libertate este

$$g(t)=\frac{\Gamma\left(\frac{s+1}{2}\right)}{\sqrt{\pi s}\,\Gamma\left(\frac{s}{2}\right)}\left(1+\frac{t^2}{s}\right)^{-\frac{s+1}{2}}.$$

In cazul acestei distribuții observăm că momentele obișnuite coincid cu momentele centrate, deoarece momentul de ordinul întîi este egal cu zero.

Avem, evident,

$$m_{2r+1}=\frac{\Gamma\left(\frac{s+1}{2}\right)}{\sqrt{\pi s}\cdot\Gamma\left(\frac{s}{2}\right)}\int_{-\infty}^{+\infty}t^{2r+1}\left(1+\frac{t^2}{s}\right)^{-\frac{s+1}{2}}dt=0,$$

deoarece funcția $t^{2r+1}\left(1+\frac{t^2}{s}\right)^{-\frac{s+1}{2}}$ este impară și

$$m_{2r}=\frac{\Gamma\left(\frac{s+1}{2}\right)}{\sqrt{\pi s}\cdot\Gamma\left(\frac{s}{2}\right)}\int_{-\infty}^{+\infty}t^{2r}\left(1+\frac{t^2}{s}\right)^{-\frac{s+1}{2}}dt.$$

Folosind substituția $\frac{t^2}{2}=s$, momentul centrat de ordinul $2r$ este

$$m_{2r}=\frac{s^r\Gamma\left(\frac{s+1}{2}\right)}{2\sqrt{\pi}\cdot\Gamma\left(\frac{s}{2}\right)}\int_{-\infty}^{+\infty}x^{r-\frac{1}{2}}(1+x)^{-\frac{s+1}{2}}dx. \qquad \text{(XI.10,1)}$$

Se știe, însă, că

$$\int_0^{\infty}x^{p-1}(1+x)^{-(p+q)}dx=\frac{\Gamma(p)\,\Gamma(q)}{\Gamma(p+q)}.$$

Atunci, rezultă că

$$\int_{-\infty}^{+\infty} x^{r-\frac{1}{2}} (1+x)^{-\frac{s+1}{2}} dx = 2 \frac{\Gamma\left(r+\frac{1}{2}\right) \cdot \Gamma\left(\frac{s}{2}-r\right)}{\Gamma\left(\frac{s+1}{2}\right)}. \quad \text{(XI.10,2)}$$

Avînd în vedere (XI.10,2), expresia (XI.10,1) devine

$$m_{2r} = \frac{s^r \Gamma\left(\frac{s+1}{2}\right)}{2\sqrt{\pi}\,\Gamma\left(\frac{s}{2}\right)} \cdot \frac{2\Gamma\left(r+\frac{1}{2}\right) \cdot \left(\frac{s}{2}-r\right)}{\Gamma\left(\frac{s+1}{2}\right)} =$$

$$= \frac{s^r}{\sqrt{\pi}} \cdot \frac{\Gamma\left(r+\frac{1}{2}\right)\Gamma\left(\frac{s}{2}-r\right)}{\Gamma\left(\frac{s}{2}\right)}. \quad \text{(XI.10,3)}$$

Observăm că momentul centrat m_{2r} există numai dacă $r < \frac{s}{2}$.

Expresia (XI.10,3) a momentului centrat de ordinul $2r$ se poate pune sub o formă mai simplă.

Intr-adevăr, folosind o proprietate cunoscută a funcției $\Gamma(x)$, avem

$$\Gamma\left(\frac{s}{2}\right) = \left(\frac{s}{2}-1\right)\left(\frac{s}{2}-2\right)\ldots\left(\frac{s}{2}-r\right)\Gamma\left(\frac{s}{2}-r\right).$$

La fel, după aceeași formulă

$$\Gamma\left(r+\frac{1}{2}\right) = \left(r-\frac{1}{2}\right)\left(r-\frac{3}{2}\right)\ldots\frac{1}{2}\cdot\Gamma\left(\frac{1}{2}\right).$$

Inlocuind în (XI.10,3) valorile de mai sus ale lui $\Gamma\left(\frac{s}{2}\right)$ și $\Gamma\left(r+\frac{1}{2}\right)$, obținem

$$m_{2r} = \frac{s^r}{\sqrt{\pi}} \cdot \frac{\left(r-\frac{1}{2}\right)\left(r-\frac{3}{2}\right)\ldots\frac{1}{2}\sqrt{\pi}\,\Gamma\left(\frac{s}{2}-r\right)}{\left(\frac{s}{2}-1\right)\left(\frac{s}{2}-2\right)\ldots\left(\frac{s}{2}-r\right)\Gamma\left(\frac{s}{2}-r\right)} =$$

$$= \frac{s^r \cdot 1 \cdot 3 \ldots (2r-3)(2r-1)}{(s-2)(s-4)\ldots(s-2r)}. \quad \text{(XI.10,4)}$$

Din (XI.10,4), particularizînd pe r, deducem momentele centrate de ordin par ale distribuției Student

$$m_2 = \frac{s}{s-2},$$

$$m_4 = \frac{3s^2}{(s-2)(s-4)},$$

$$m_6 = \frac{15s^3}{(s-2)(s-4)(s-6)},$$

. . . . . .

Teoremă. *Dacă s, numărul gradelor de libertate, crește nemărginit, atunci distribuția Student cu s grade de libertate tinde către distribuția normală $N(0, 1)$.*

Pentru demonstrarea teoremei, este suficient să arătăm că șirul de momente m_p, oricare ar fi p întreg și pozitiv, tinde către momentele distribuției normale $N(0, 1)$, dacă $s \to \infty$.

In acest caz, funcțiile de repartiție corespunzătoare tind, de asemenea, către funcția de repartiție a distribuției normale $N(0, 1)$.

Dacă p este impar, atunci momentele distribuției Student sînt, la limită, egale cu momentele distribuției normale.

Dacă p este, însă, par, obținem

$$\lim_{s\to\infty} m_{2r} = \lim_{s\to\infty} \frac{s^r \cdot 1 \cdot 3 \ldots (2r-3)(2r-1)}{(s-2)(s-4)\ldots(s-2r)} = 1 \cdot 3 \cdot 5 \ldots (2r-1),$$

întrucît numărătorul și numitorul sînt polinoame în s de gradul r cu coeficienții respectiv egali cu $1 \cdot 3 \cdot 5 \ldots (2r-1)$ și 1.

Cum, însă, $1 \cdot 3 \cdot 5 \ldots (2r-1)$ este momentul de ordinul $2r$ al distribuției normale $N(0, 1)$, teorema este demonstrată.

§ 11. Alte variabile care dau naștere la distribuția Student sau la distribuții înrudite cu ea

Să considerăm variabilele următoare

$$y = \sqrt{s(s-1)} \cdot \frac{\bar{x}}{\sqrt{\sum_{i=1}^{s} (x_i - \bar{x})^2}} \qquad \text{(XI.11,1)}$$

şi

$$z = \frac{x_s}{\sqrt{\frac{x_1^2+x_2^2+\ldots+x_s^2}{s}}}, \qquad \text{(XI.11,2)}$$

unde $x_i\,(i=1,2,\ldots,s)$ sînt variabile normale $N(0,\sigma)$, iar $\bar{x} = \frac{x_1+x_2+\ldots+x_s}{s}$, adică media de selecţie.

In cazul de faţă, ne propunem să determinăm legea de probabilitate a distribuţiilor variabilelor considerate.

1) Pentru determinarea legii de probabilitate vom presupune că $x_1, x_2, \ldots, x_s$ sînt coordonatele unui punct $M(x_1, x_2, \ldots, x_s)$ dintr-un spaţiu euclidian cu s dimensiuni.

Facem o schimbare de axe de coordonate cu ajutorul transformării ortogonale de mai jos

$$\begin{aligned}
y_1 &= a_{11}x_1 + a_{12}x_2 + \ldots + a_{1s}x_s,\\
y_2 &= a_{21}x_1 + a_{22}x_2 + \ldots + a_{2s}x_s,\\
&\ldots\ldots\ldots\ldots\ldots\ldots\\
y_{s-1} &= a_{s-1,1}x_1 + a_{s-1,2}x_2 + \ldots + a_{s-1,s}x_s,\\
y_s &= \frac{1}{\sqrt{s}}x_1 + \frac{1}{\sqrt{s}}x_2 + \ldots + \frac{1}{\sqrt{s}}x_s.
\end{aligned} \qquad \text{(XI.11,3)}$$

Transformarea (XI.11,3) fiind ortogonală, lasă invariantă, în noile axe de coordonate, expresia

$$x_1^2 + x_2^2 + \ldots + x_s^2,$$

iar variabilele $x_i\,(i=1,2,\ldots,n)$ fiind independente, şi variabilele $y_i\,(i=1,2,\ldots,n)$ sînt independente.

Prin urmare,

$$y_1^2 + y_2^2 + \ldots + y_s^2 = x_1^2 + x_2^2 + \ldots + x_s^2. \qquad \text{(XI.11,4)}$$

Din aceeaşi transformare ortogonală, mai rezultă

$$y_s^2 = \frac{(x_1+x_2+\ldots+x_s)^2}{s} = s\bar{x}^2,$$

de unde, deducem

$$\bar{x} = \frac{y_s}{\sqrt{s}} \qquad \text{(XI.11,5)}$$

Scanned with OKEN Scanner

şi deci

$$y_1^2+y_2^2+\ldots+y_{s-1}^2=\sum_{i=1}^{s}x_i^2-s\bar{x}^2=\sum_{i=1}^{s}(x_i-\bar{x})^2. \quad \text{(XI.11,6)}$$

Atunci, ţinînd seama de relaţiile (XI.11,5) şi (XI.11,6), variabila (XI.11,1) se va scrie

$$y=\sqrt{s(s-1)}\cdot\frac{\bar{x}}{\sqrt{\sum_{i=1}^{s}(x_i-\bar{x})^2}}=\frac{y_s}{\sqrt{\frac{y_1^2+y_2^2+\ldots+y_{s-1}^2}{s-1}}}.$$

Se ştie din cele stabilite mai înainte că variabila

$$\frac{y_s}{\sqrt{\frac{y_1^2+y_2^2+\ldots+y_{s-1}^2}{s-1}}}$$

dă naştere unei distribuţii Student cu $s-1$ grade de libertate, care este astfel şi distribuţia variabilei (XI.11,1).

Cu alte cuvinte, *media de selecţie, raportată la rădăcina pătrată a dispersiei de selecţie, urmează o lege Student cu* $s-1$ *grade de libertate.*

2) Să determinăm, acum, legea de probabilitate a variabilei (XI.11,2). Este evident că, dacă notăm

$$X=x_s \quad \text{şi} \quad Y=\sqrt{\frac{x_1^2+x_2^2+\ldots+x_s^2}{s}},$$

X şi Y sînt două variabile dependente.

Deci, în acest caz, nu vom mai putea pleca de la probabilitatea elementară pentru punctul (X, Y), aşa cum am procedat cînd am considerat variabila

$$\frac{x_{s+1}}{\sqrt{\frac{x_1^2+x_2^2+\ldots+x_s^2}{s}}},$$

unde x_{s+1} şi $\sqrt{\frac{\sum_{i=1}^{s}x_i^2}{s}}$ erau variabile independente.

De aceea, vom urma aceeaşi cale ca mai sus, adică vom considera că $x_1, x_2, \ldots, x_s$ sînt coordonatele unui punct $M(x_1, x_2, \ldots, x_s)$ în spaţiul euclidian cu s dimensiuni.

Trecem de la vechiul sistem de axe la un nou sistem de axe, prin transformarea următoare de variabile

$$\begin{aligned} x_1 &= \rho \cos\theta_1 \cos\theta_2 \ldots \cos\theta_{s-2} \cos\theta_{s-1}, \\ x_2 &= \rho \cos\theta_1 \cos\theta_2 \ldots \cos\theta_{s-2} \sin\theta_{s-1}, \\ &\cdots\cdots\cdots\cdots\cdots\cdots \\ x_{s-1} &= \rho \cos\theta_1 \sin\theta_2, \\ x_s &= \rho \sin\theta_1. \end{aligned} \qquad \text{(XI.11,7)}$$

Din transformarea (XI.11,7), deducem uşor

$$Z = \frac{x_s}{\sqrt{\frac{x_1^2 + x_2^2 + \ldots + x_s^2}{s}}} = \sqrt{s} \cdot \sin\theta_1. \qquad \text{(XI.11,8)}$$

Atunci, legea de probabilitate a variabilei Z va rezulta imediat cunoscînd legea de probabilitate a lui θ_1.

Insă, legea de probabilitate elementară a variabilei θ_1 cu s grade de libertate a fost determinată anterior (XI.9,1) şi

$$\frac{1}{\sqrt{\pi}} \cdot \frac{\Gamma\left(\frac{s+1}{2}\right)}{\Gamma\left(\frac{s}{2}\right)} \cos^{s-1}\theta_1 d\theta_1$$

este legea de probabilitate

$$\frac{1}{\sqrt{\pi}} \cdot \frac{\Gamma\left(\frac{s}{2}\right)}{\Gamma\left(\frac{s-1}{2}\right)} \cos^{s-2}\theta_1 d\theta_1. \qquad \text{(XI.11,9)}$$

In baza relaţiei (XI.11,8) avem

$$\cos\theta_1 = \sqrt{1 - \frac{z^2}{s}}$$

și

$$d\theta_1 = \frac{dz}{\sqrt{s}\cdot\sqrt{1-\frac{z^2}{s}}}.$$

Ducînd, în sfîrșit, valorile lui $\cos\theta_1$ și $d\theta_1$ în (XI.11,9), obținem legea de probabilitate căutată a variabilei (XI.11,2)

$$\frac{\Gamma\left(\frac{s}{2}\right)}{\sqrt{\pi s}\,\Gamma\left(\frac{s-1}{2}\right)}\cdot\left(1-\frac{z^2}{s}\right)^{\frac{s-3}{2}}. \qquad \text{(XI.11,10)}$$

In continuare, examinăm reprezentarea grafică a legii de probabilitate (XI.11,10) pentru cazurile cînd s ia valorile 2, 3 și 4.

a) $s=2$. Legea de probabilitate (XI.11,10) devine

$$g_2(x) = \frac{1}{\pi\sqrt{2}}\left(1-\frac{x^2}{2}\right)^{-\frac{1}{2}}, \qquad \text{(XI.11,11)}$$

unde am înlocuit variabila z prin x.

Se vede că intervalul de existență al funcției $g_2(x)$ este $(-\sqrt{2}, \sqrt{2})$, iar în acest interval ea are numai valori pozitive.

Pentru $x=\pm\sqrt{2}$, funcția devine infinită.

Derivînd, obținem

$$g_2'(x) = \frac{x}{2\left(1-\frac{x^2}{2}\right)}\,g(x)\begin{cases} <0, & -\sqrt{2}<x<0, \\ >0, & 0<x<\sqrt{2}. \end{cases}$$

Rezultă că, pentru $x=0$, funcția (XI.11,11) are un minim egal cu $\frac{1}{\pi\sqrt{2}}$.

Ea este descrescătoare în intervalul $(-\sqrt{2}, 0)$ și crescătoare în intervalul $(0, \sqrt{2})$.

b) $s=3$. Legea de probabilitate (XI.11,10) se scrie

$$g_3(x) = \frac{1}{2\sqrt{3}}. \qquad \text{(XI.11,12)}$$

Scanned with OKEN Scanner

In acest caz, funcţia (XI.11,12) este o dreaptă paralelă cu axa absciselor. După cum vom vedea mai departe, aceasta poartă numele de *distribuţia rectangulară*.

c) $s=4$. Legea de probabilitate (XI.11,10) este

$$g_4(x)=\frac{1}{\pi}\left(1-\frac{x^2}{4}\right)^{\frac{1}{2}}. \qquad \text{(XI.11,13)}$$

Intervalul de existenţă este, evident, $(-2, 2)$, în care funcţia ia, de asemenea, numai valori pozitive.

Dacă $x=\pm 2$, funcţia $g_4(x)$ se anulează.

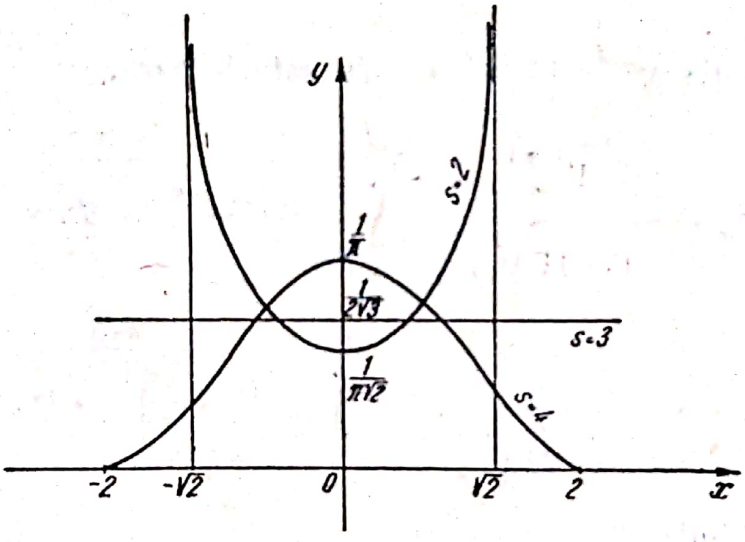


Fig. XI.11,1

Luînd derivata, găsim

$$g_4'(x)=-\frac{x}{4\left(1-\frac{x^2}{4}\right)}\cdot g(x) \qquad \begin{cases} >0, & -2<x<0, \\ <0, & 0<x<2. \end{cases}$$

De unde deducem că, pentru $x=0$, funcţia (XI.11,13) are un maxim egal cu $\frac{1}{\pi}$, fiind crescătoare în intervalul $(-2, 0)$ şi descrescătoare în intervalul $(0, 2)$.

Pentru cele trei cazuri examinate, arătăm în fig. XI.11,1 graficele funcţiilor considerate.

După reprezentarea grafică, să trecem acum la calculul momentelor distribuţiei cu legea de probabilitate (XI.11,13).

Scanned with OKEN Scanner

Această lege de probabilitate, fiind simetrică în raport cu axa ordonatelor, momentele obişnuite vor coincide cu momentele centrate. În acelaşi timp, am văzut că distribuţia există numai pentru valorile lui x cuprinse între $-\sqrt{s}$ şi $\sqrt{s}$.
Prin urmare,

$$m_{2r+1}=\frac{\Gamma\left(\frac{s}{2}\right)}{\sqrt{\pi s}\,\Gamma\left(\frac{s-1}{2}\right)}\int_{-\sqrt{s}}^{\sqrt{s}} x^{2r+1}\left(1-\frac{x^2}{s}\right)^{\frac{s-3}{2}}dx=0,$$

deoarece funcţia $x^{2r+1}\cdot\left(1-\frac{x^2}{s}\right)^{\frac{s-3}{2}}dx$ este impară,

$$m_{2r}=\frac{\Gamma\left(\frac{s}{2}\right)}{\Gamma(\pi s)\,\Gamma\left(\frac{s-1}{2}\right)}\int_{-\sqrt{s}}^{\sqrt{s}} x^{2r}\left(1-\frac{x^2}{s}\right)^{\frac{s-3}{2}}dx=$$

$$=\frac{2\Gamma\left(\frac{s}{2}\right)}{\sqrt{\pi s}\cdot\Gamma\left(\frac{s-1}{2}\right)}\cdot\int_{0}^{\sqrt{s}} x^{r}\left(1-\frac{x^2}{s}\right)^{\frac{s-3}{2}}\cdot dx. \qquad \text{(XI.11,14)}$$

Făcînd substituţia

$$x^2=\frac{sy}{1+y},$$

expresia (XI.11,14) a momentului de ordinul $2r$ devine

$$m_{2r}=\frac{s^r\Gamma\left(\frac{s}{2}\right)}{\sqrt{\pi}\,\Gamma\left(\frac{s-1}{2}\right)}\int_{0}^{\infty} y^{r-\frac{1}{2}}(1+y)^{-\left(r+\frac{s}{2}\right)}\cdot dy.$$

Integrala obţinută fiind de tipul

$$\int_{0}^{\infty} y^{p-1}(1+y)^{-(p+q)}dy=\frac{\Gamma(p)\cdot\Gamma(q)}{\Gamma(p+q)},$$

putem încă scrie

$$m_{2r}=\frac{s^r\Gamma\left(\frac{s}{r}\right)}{\sqrt{\pi}\cdot\Gamma\left(\frac{s-1}{2}\right)}\cdot\frac{\Gamma\left(r+\frac{1}{2}\right)\cdot\Gamma\left(\frac{s-1}{2}\right)}{\Gamma\left(r+\frac{s}{2}\right)}=\frac{s^r\cdot\Gamma\left(\frac{s}{2}\right)\cdot\Gamma\left(r+\frac{1}{2}\right)}{\sqrt{\pi}\cdot\Gamma\left(r+\frac{s}{2}\right)}\cdot\quad\text{(XI.11,15)}$$

Introducînd în (XI.11,15) dezvoltările factoriale

$$\Gamma(r+1)=\left(r-\frac{1}{2}\right)\left(r-\frac{3}{2}\right)\ldots\frac{1}{2}\Gamma\left(\frac{1}{2}\right)$$

şi

$$\Gamma\left(r+\frac{s}{2}\right)=\left(r+\frac{s}{2}-1\right)\left(r+\frac{s}{2}-2\right)\ldots\frac{s}{2}\Gamma\left(\frac{s}{2}\right),$$

obţinem

$$m_{2r}=\frac{s^r\left(r-\frac{1}{2}\right)\left(r-\frac{3}{2}\right)\ldots\frac{1}{2}\Gamma\left(\frac{s}{2}\right)}{\sqrt{\pi}\cdot\left(r+\frac{s}{2}-1\right)\left(r+\frac{s}{2}-2\right)\ldots\frac{s}{2}}=s^r.$$

In particular, dacă $r=1, 2, \ldots$, rezultă imediat primele momente pare ale distribuţiei

$$m_2=1,$$

$$m_4=\frac{3s^2}{s(s+2)},$$

$$m_6=\frac{15s^3}{s(s+2)(s+4)},$$

. . . . . .

Este interesant de observat că dispersia m_2 este independentă de s numărul gradelor de libertate.

De asemenea, se constată uşor că şi această distribuţie tinde către distribuţia normală $N(0, 1)$, cînd s creşte nemărginit.

In adevăr, avem

$$\lim_{s\to\infty}m_{2r}=\lim_{s\to\infty}\frac{s^r\cdot1\cdot3\ldots(2r-1)}{s(s+2)\ldots(s+2r-2)}=1\cdot3\ldots(2r-1)$$

şi

$$m_{2r+1}=0.$$

Intrucît momentele distribuţiei în cauză tind către momentele distribuţiei normale $N(0, 1)$, atunci şi funcţiile de repartiţie corespunzătoare vor tinde către funcţia de repartiţie a aceleiaşi distribuţii normale $N(0, 1)$.

Această proprietate de limită se poate bănui de altfel şi din reprezentarea grafică a distribuţiei, unde se vede că pentru $s=4$, curba se apropie de forma curbei normale.

§ 12. Distribuţia Snedecor

Fie două variabile x şi y, care urmează fiecare, respectiv, legile de probabilitate $H(s_1, \sigma)$ şi $H(s_2, \sigma)$.

Ne propunem să determinăm legea de probabilitate, adică funcţia de frecvenţă a variabilei dată de raportul

$$u = \frac{\frac{x}{s_1}}{\frac{y}{s_2}} = \frac{s_2 x}{s_1 y}. \qquad \text{(XI.12,1)}$$

1. Variabilele x şi y fiind independente, rezultă că probabilitatea elementară pentru punctul (x, y) este

$$\frac{1}{2^{\frac{s_1+s_2}{2}} \cdot \sigma^{s_1+s_2} \cdot \Gamma\left(\frac{s_1}{2}\right) \cdot \Gamma\left(\frac{s_2}{2}\right)} \cdot x^{\frac{s_1}{2}-1} y^{\frac{s_2}{2}-1} \cdot e^{-\frac{x+y}{2\sigma^2}} \cdot dxdy. \qquad \text{(XI.12,2)}$$

Prin substituţia

$$x = \frac{s_1}{s_2} uv, \quad y = v,$$

această probabilitate devine

$$\left(\frac{s_1}{s_2}\right)^{\frac{s_1}{2}} \frac{1}{2^{\frac{s_1+s_2}{2}} \cdot \sigma^{s_1+s_2} \cdot \Gamma\left(\frac{s_1}{2}\right) \cdot \Gamma\left(\frac{s_2}{2}\right)} \cdot u^{\frac{s_1}{2}-1} \cdot v^{\frac{s_1+s_2}{2}-1} \cdot e^{-\frac{x+y}{2\sigma^2}} \cdot dudv. \qquad \text{(XI.12,3)}$$

Pentru a găsi legea de probabilitate a variabilei u, integrăm în raport cu variabila v între limitele 0 şi ∞.

Astfel, vom avea

$$h(u,\ s_1,\ s_2)\,du=\left(\frac{s_1}{s_2}\right)^{\frac{s_1}{2}}\frac{u^{\frac{s_1}{2}-1}\,du}{2^{\frac{s_1+1}{2}}\sigma^{s_1+s_2}\Gamma\left(\frac{s_1}{2}\right)\Gamma\left(\frac{s_2}{2}\right)}\int_0^{\infty}v^{\frac{s_1+s_2}{2}-1}e^{-\frac{v}{2\sigma^2}\left(1+\frac{s_1}{s_2}u\right)}dv.$$

De unde,

$$h(u,\ s_1,\ s_2)=\left(\frac{s_1}{s_2}\right)^{\frac{s_1}{2}}\frac{u^{\frac{s_1}{2}-1}}{2^{\frac{s_1+s_2}{2}}\sigma^{s_1+s_2}\Gamma\left(\frac{s_1}{2}\right)\Gamma\left(\frac{s_2}{2}\right)}\cdot\frac{\Gamma\left(\frac{s_1+s_2}{2}\right)}{\left(\frac{1+\frac{s_1}{s_2}u}{2\sigma^2}\right)^{\frac{s_1+s_2}{2}}}=$$

$$=\left(\frac{s_1}{s_2}\right)^{\frac{s_1}{2}}\frac{\Gamma\left(\frac{s_1+s_2}{2}\right)}{\Gamma\left(\frac{s_1}{2}\right)\Gamma\left(\frac{s_2}{2}\right)}\cdot u^{\frac{s_1}{2}-1}\left(1+\frac{s_1}{s_2}u\right)^{-\frac{s_1+s_2}{2}},\qquad\text{(XI.12,4)}$$

deoarece $\int_0^{\infty}v^{\frac{s_1+s_2}{2}}e^{-\frac{v}{2\sigma^2}\left(1+\frac{s_1}{s_2}u\right)}dv$ este de tipul $\Gamma(x)$.

Funcția $h(u,\ s_1,\ s_2)$ este *funcția de frecvență a distribuției Snedecor cu s_1 și s_2 grade de libertate*, care, după cum se vede, este independentă de parametrul σ.

2. Vom arăta, acum, o altă metodă pentru găsirea funcției de frecvență (XI.12,4) care prezintă interes prin faptul că dă o interpretare geometrică rezultatului obținut.

Variabilele x și y, urmînd fiecare o lege de probabilitate Helmert-Pearson, reprezintă cîte o sumă de s_1 și s_2 pătrate, respectiv, adică

$$x=\sum_{i=1}^{s_1}x_i^2\quad\text{și}\quad y=\sum_{i=1}^{s_2}y_j^2,$$

unde x_i și y_j sînt variabile normale $N(0,\ \sigma)$, independente.

Atunci, expresia (XI.12,1) a variabilei u se va scrie

$$u=\frac{s_2}{s_1}\cdot\frac{x_1^2+x_2^2+\ldots+x_{s_1}^2}{y_1^2+y_2^2+\ldots+y_{s_2}^2}.\qquad\text{(XI.12,5)}$$

Presupunînd că $x_1, x_2, \ldots, x_{s_1}, y_1, y_2, \ldots, y_{s_2}$ sînt coordonatele unui punct $M(x_1, x_2, \ldots, x_{s_1}, y_1, y_2, \ldots, y_{s_2})$ într-un spaţiu euclidian cu $s_1 + s_2$ dimensiuni, probabilitatea elementară pentru acest punct va fi

$$\frac{1}{(2\pi)^{\frac{s_1+s_2}{2}}\sigma^{s_1+s_2}} e^{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{s_1} x_i^2 + \sum_{j=1}^{s_2} y_j^2\right)} dx_1 \cdot dx_2 \ldots dx_{s_1} \cdot dy_1 \ldots dy_{s_2}. \quad \text{(XI.12,6)}$$

Pentru calcularea probabilităţii elementare (XI.12,6), proiectăm punctul $M(x_1, x_2, \ldots, x_{s_1}, y_1, y_2, \ldots, y_{s_2})$ pe subspaţiile liniare

$$L_1(x_1, x_2, \ldots, x_{s_1}, 0, 0, \ldots, 0)$$

şi

$$L_2(0, 0, \ldots, 0, y_1, y_2, \ldots, y_{s_2}).$$

Efectuînd în fiecare dintre aceste subspaţii liniare o schimbare de variabile în coordonate polare

$$\begin{aligned}
x_1 &= \rho \cos\theta \cos\varphi_1 \ldots \cos\varphi_{s_1-2} \cos\varphi_{s_1-1}, \\
x_2 &= \rho \cos\theta \cos\varphi_1 \ldots \cos\varphi_{s_1-2} \sin\varphi_{s_1-1}, \\
&\ldots\ldots\ldots\ldots\ldots\ldots \\
x_{s_1} &= \rho \cos\theta \sin\varphi_1; \qquad \text{(XI.12,7)} \\
y_1 &= \rho \sin\theta \cos\psi_1 \ldots \cos\psi_{s_2-2} \cos\psi_{s_2-1}, \\
y_2 &= \rho \sin\theta \cos\psi_1 \ldots \cos\psi_{s_2-2} \sin\psi_{s_2-1}, \\
&\ldots\ldots\ldots\ldots\ldots\ldots \\
y_{s_2} &= \rho \sin\theta \sin\psi_1,
\end{aligned}$$

din transformarea (XI.12,7), rezultă relaţiile

$$x_1^2 + x_2^2 + \ldots + x_{s_1}^2 = \rho^2 \cos^2\theta$$

şi

$$y_1^2 + y_2^2 + \ldots + y_{s_2}^2 = \rho^2 \sin^2\theta$$

care ne permit să scriem variabila (XI.12,5) sub forma

$$u = \frac{s_2}{s_1} \operatorname{ctg}^2\theta. \qquad \text{(XI.12,8)}$$

Scanned with OKEN Scanner

Variabila lui Snedecor se exprimă numai în raport cu unghiul θ din formulele (XI.12,7) şi va fi suficient să găsim funcţia de frecvenţă a lui θ pentru a putea determina funcţia de frecvenţă a lui u. Determinantul funcţional al transformării de variabile (XI.12,7) se calculează uşor, dacă considerăm mai întîi schimbarea de variabile

$$\begin{aligned}
x_1 &= \rho_1 \cos\bar{\varphi}_1 \ldots \cos\bar{\varphi}_{s_1-2} \cos\bar{\varphi}_{s_1-1},\\
x_2 &= \rho_1 \cos\bar{\varphi}_1 \ldots \cos\bar{\varphi}_{s_1-2} \sin\bar{\varphi}_{s_1-1},\\
&\ldots\ldots\ldots\ldots\ldots\ldots\\
x_{s_1} &= \rho_1 \sin\bar{\varphi}_1;\\
y_1 &= \rho_2 \cos\bar{\psi}_1 \ldots \cos\bar{\psi}_{s_2-2} \cos\bar{\psi}_{s_2-1},\\
y_2 &= \rho_2 \cos\bar{\psi}_1 \ldots \cos\bar{\psi}_{s_2-2} \sin\bar{\psi}_{s_2-1},\\
&\ldots\ldots\ldots\ldots\ldots\ldots\\
y_{s_2} &= \rho_2 \sin\bar{\psi}_1.
\end{aligned} \qquad \text{(XI.12,9)}$$

Determinantul funcţional al unei transformări de forma (XI.12,9) a mai fost calculat anterior pentru cazul a s şi $s+1$ variabile.

In cazul de faţă, avînd $s_1 + s_2$ variabile $(\rho_1, \bar{\varphi}_1, \ldots, \bar{\varphi}_{s_1-1}, \rho_2, \bar{\psi}_1, \ldots, \bar{\psi}_{s_2-1})$ determinantul funcţional al transformării (XI.12,9) se scrie

$$\rho_1^{s_1-1}\rho_2^{s_2-1}\cos^{s_1-2}\bar{\varphi}_1 \ldots \cos\bar{\varphi}_{s_1-2}\cos^{s_2-2}\bar{\psi}_1 \ldots \cos\bar{\psi}_{s_2-2}. \qquad \text{(XI.12,10)}$$

Printr-o nouă substituţie

$$\begin{aligned}
&\rho_1 = \rho\cos\theta,\ \bar{\varphi}_1 = \varphi_1, \ldots, \bar{\varphi}_{s_1-1} = \varphi_{s_1-1},\\
&\rho_2 = \rho\sin\theta,\ \bar{\psi}_1 = \psi_1, \ldots, \bar{\psi}_{s_2-1} = \psi_{s_2-1},
\end{aligned} \qquad \text{(XI.12,11)}$$

se trece imediat de la determinantul funcţional (XI.12,10) la determinantul funcţional al transformării (XI.12,7).

Calcule simple, pe care nu le mai reproducem aici, ne arată că determinantul funcţional al noii substituţii (XI.12,11) este egal cu ρ şi, deci, obţinem următoarea expresie pentru determinantul funcţional căutat al transformării (XI.12,7)

$$\rho^{s_1+s_2-1}\cos^{s_1-1}\theta\sin^{s_2-1}\theta \cdot \cos^{s_1-2}\varphi_1 \ldots \cos\varphi_{s_1-2}\cos^{s_2-2}\psi_1 \ldots \cos\psi_{s_2-2}.$$

De unde deducem, în urma transformării de variabile (XI.12,7), că probabilitatea elementară (XI.12,6) a punctului $M(x_1, x_2, \ldots, x_{s_1}, y_1, \ldots, y_{s_2})$ este

$$\frac{1}{(2\pi)^{\frac{s_1+s_2}{2}}\sigma^{s_1+s_2}} \cdot e^{-\frac{\rho^2}{2\sigma^2}} \cdot \rho^{s_1+s_2-1} \left| \cos^{s_1-1}\theta \cdot \sin^{s_2-1}\theta \cdot \cos^{s_2-2}\varphi_1 \ldots \right.$$

$$\left. \ldots \cos\varphi_{s_1-2} \cdot \cos^{s_2-2}\psi_1 \ldots \cos\psi_{s_2-2} \right| d\rho \cdot d\theta \cdot d\varphi_1 \ldots d\psi_{s_1-1}, \quad \text{(XI.12,12)}$$

unde am efectuat înlocuirile $\sum_{i=1}^{s_1} x_i^2 = \rho^2 \cos^2\theta$ și $\sum_{j=1}^{s_2} y_j^2 = \rho^2 \sin^2\theta$.

Vom găsi legea de probabilitate a variabilei θ, integrînd expresia (XI.12,12) în raport cu variabilele $\rho, \varphi_1, \ldots, \varphi_{s_1-1}, \psi_1, \ldots, \psi_{s_2-1}$. Vom avea

$$2^{s_1+s_2-1} \frac{\cos^{s_1-1}\theta \cdot \sin^{s_2-1}\theta \cdot d\theta}{(2\pi)^{\frac{s_1+s_2}{2}} \cdot \sigma^{s_1+s_2}} \cdot \int_0^{\infty} \rho^{s_1+s_2-1} \cdot e^{-\frac{\rho^2}{2\sigma^2}} \cdot d\rho \int_0^{\frac{\pi}{2}} \cos^{s_1-2}\varphi_1 d\varphi_1 \ldots$$

$$\ldots \int_0^{\frac{\pi}{2}} d\varphi_{s_1-1} \int_0^{\frac{\pi}{2}} \cos^{s_2-2}\psi_1 d\psi_1 \ldots \int_0^{\frac{\pi}{2}} d\psi_{s_2-2} =$$

$$= A \cos^{s_1-1}\theta \sin^{s_2-1}\theta d\theta \quad \left(0 \leqq \theta \leqq \frac{\pi}{2}\right), \quad \text{(XI.12,13)}$$

unde

$$A = \frac{1}{\int_0^{\frac{\pi}{2}} \cos^{s_1-1}\theta \sin^{s_2-1}\theta d\theta} = 2\frac{\Gamma\left(\frac{s_1+s_2}{2}\right)}{\Gamma\left(\frac{s_1}{2}\right)\Gamma\left(\frac{s_2}{2}\right)}.$$

Legea de probabilitate elementară a lui θ este, prin urmare,

$$\frac{2\Gamma\left(\frac{s_1+s_2}{2}\right)}{\Gamma\left(\frac{s_1}{2}\right) \cdot \Gamma\left(\frac{s_2}{2}\right)} \cos^{s_1-1}\theta \cdot \sin^{s_2-1}\theta \cdot d\theta \quad \left(0 \leqq \theta \leqq \frac{\pi}{2}\right). \quad \text{(XI.12,14)}$$

Scanned with OKEN Scanner

De la această lege, putem trece uşor la legea de probabilitate a variabilei (XI.12,1).

In adevăr, din relaţia (XI.12,8), deducem

$$\cos\theta = \left(\frac{s_1}{s_2}\right)^{\frac{1}{2}} \frac{u^{\frac{1}{2}}}{\left(1+\frac{s_1}{s_2}u\right)^{\frac{1}{2}}},$$

$$\sin\theta = \frac{1}{\left(1+\frac{s_1}{s_2}u\right)^{\frac{1}{2}}}, \qquad \text{(XI.12,15)}$$

$$d\theta = \frac{1}{2}\left(\frac{s_1}{s_2}\right)^{\frac{1}{2}} u^{-\frac{1}{2}}\left(1+\frac{s_1}{s_2}u\right)^{-1}\cdot du,$$

deoarece $0<\theta<\frac{\pi}{2}$.

Am luat valoarea pozitivă a lui $d\theta$, deoarece probabilitatea este o funcţie pozitivă, iar $\sin\theta>0$, $\cos\theta>0$.

In sfîrşit, introducînd în (XI.12,14) valorile lui $\cos\theta$, $\sin\theta$ şi $d\theta$ date de relaţiile (XI.12,15) regăsim legea de probabilitate (XI.12,4) a variabilei u

$$h(u,\ s_1,\ s_2) = \left(\frac{s_1}{s_2}\right)^{\frac{s_1}{2}} \cdot \frac{\Gamma\left(\frac{s_1+s_2}{2}\right)}{\Gamma\left(\frac{s_1}{2}\right)\cdot\Gamma\left(\frac{s_2}{2}\right)} \cdot u^{\frac{s_1}{2}-1}\left(1+\frac{s_1}{s_2}u\right)^{-\frac{s_1+s_2}{2}} \quad (0\leqslant u\leqslant\infty).$$

In legătură cu distribuţia Snedecor, facem următoarea observaţie. Am arătat la timp că distribuţia Student cu s grade de libertate corespunde variabilei (XI.9,5)

$$t = \sqrt{s}\cdot\frac{x_{s+1}}{\sqrt{x_1^2+x_2^2+\ldots+x_s^2}},$$

unde $x_i\,(i=1,\ 2,\ \ldots,\ s+1)$ sînt variabile independente, normale $N(0,\ \sigma)$.

Ridicînd la pătrat, rezultă

$$t^2=\frac{s}{1}\cdot\frac{x_{s+1}^2}{x_1^2+x_2^2+\ldots+x_s^2}=u.$$

Variabila $t^2=u$ are, după cele stabilite mai înainte, o distribuţie Snedecor cu 1 şi $s+1$ grade de libertate cu legea de probabilitate $h(u, 1, s)$.

Cu alte cuvinte, *pătratul variabilei t, care urmează o lege de probabilitate Student* $g(t, s)$ *cu s grade de libertate, urmează, la rîndul său, o lege de probabilitate Snedecor* $h(u, 1, s)$ *cu 1 şi s grade de libertate.*

Momentele distribuţiei Snedecor se calculează direct. Astfel, momentul de ordinul r, potrivit definiţiei sale, este

$$M_r=\left(\frac{s_1}{s_2}\right)^{\frac{s_1}{2}}\cdot\frac{\Gamma\left(\frac{s_1+s_2}{2}\right)}{\Gamma\left(\frac{s_1}{2}\right)\Gamma\left(\frac{s_2}{2}\right)}\int_0^\infty u^r\cdot u^{\frac{s_1}{2}-1}\left(1+\frac{s_1}{s_2}u\right)^{-\frac{s_1+s_2}{2}}\cdot du. \qquad \text{(XI.12,16)}$$

Efectuînd substituţia $v=\frac{s_1}{s_2}u$, expresia de mai sus devine

$$M_r=\left(\frac{s_1}{s_2}\right)^{\frac{s_1}{2}}\left(\frac{s_2}{s_1}\right)^{r+\frac{s_1}{2}}\cdot\frac{\Gamma\left(\frac{s_1+s_2}{2}\right)}{\Gamma\left(\frac{s_1}{2}\right)\Gamma\left(\frac{s_2}{2}\right)}\int_0^\infty v^{r+\frac{s_1}{2}-1}(1+v)^{-\frac{s_1+s_2}{2}}dv=$$

$$=\left(\frac{s_2}{s_1}\right)^r\cdot\frac{\Gamma\left(r+\frac{s_1}{2}\right)\cdot\Gamma\left(\frac{s_2}{2}-r\right)}{\Gamma\left(\frac{s_1}{2}\right)\cdot\Gamma\left(\frac{s_2}{2}\right)}, \qquad \text{(XI.12,17)}$$

avînd în vedere că $\int_0^\infty v^{r+\frac{s_2}{2}-1}(1+v)^{-\frac{s_1+s_2}{2}}dv$ este o integrală de tipul $\int_0^\infty x^p(1+x)^{-(p+q)}dx=\frac{\Gamma(p)\cdot\Gamma(q)}{\Gamma(p+q)}$.

Este de remarcat că M_r există, numai dacă $r<\frac{s_2}{2}$.

Scanned with OKEN Scanner

Dacă vom înlocui în (XI.12,17) pe $\Gamma\left(r+\frac{s_1}{2}\right)$ şi $\Gamma\left(\frac{s_2}{2}\right)$ cu produsele factoriale

$$\Gamma\left(r+\frac{s_1}{2}\right)=\left(r+\frac{s_1}{2}-1\right)\left(r+\frac{s_1}{2}-2\right)\ldots\frac{s_1}{2}\,\Gamma\left(\frac{s_1}{2}\right)$$

şi

$$\Gamma\left(\frac{s_2}{2}\right)=\left(\frac{s_2}{2}-1\right)\left(\frac{s_2}{2}-2\right)\ldots\left(\frac{s_2}{2}-r\right)\Gamma\left(\frac{s_2}{2}-r\right),$$

obţinem

$$M_r=\frac{s_2^r}{s_1^r}\cdot\frac{\left(r+\frac{s_1}{2}-1\right)\left(r+\frac{s_1}{2}-2\right)\ldots\frac{s_1}{2}}{\left(\frac{s_2}{2}-1\right)\left(\frac{s_2}{2}-2\right)\ldots\left(\frac{s_2}{2}-r\right)}=\frac{s_2^r}{s_1^r}\cdot\frac{s_1\,(s_1+2)\ldots(s_1+2r-2)}{(s_2-2)\,(s_2-4)\ldots(s_2-2r)}\cdot$$

Dînd lui r valorile 1, 2, ..., deducem primele momente ale distribuţiei Snedecor cu s_1 şi s_2 grade de libertate

$$M_1=\frac{s_2}{s_2-2},$$

$$M_2=\frac{s_2^2}{s_1}\cdot\frac{s_1+2}{(s_2-2)\,(s_2-4)},$$

$$M_3=\frac{s_2^3}{s_1^2}\cdot\frac{(s_1+2)\,(s_1+4)}{(s_2-2)\,(s_2-4)\,(s_2-6)}\cdot$$

. . . . . .

De aici, se mai observă că *media distribuţiei Snedecor este independentă de* s_1, adică ea depinde numai de numărul s_2.

Vom încheia acest paragraf, arătînd că, prin transformarea de variabilă

$$z=\frac{1}{2}\ln u,$$

deducem din legea de probabilitate Snedecor o alta, cunoscută sub numele de *legea de probabilitate a distribuţiei* z.

In adevăr, prin transformarea de variabilă de mai sus, (XI.12,4) devine

$$h(z,\,s_1,\,s_2)=2\left(\frac{s_1}{s_2}\right)^{\frac{s_1}{2}}\cdot\frac{\Gamma\left(\frac{s_1+s_2}{2}\right)}{\Gamma\left(\frac{s_1}{2}\right)\cdot\Gamma\left(\frac{s_2}{2}\right)}\cdot e^{s_1 z}\left(1+\frac{s_1}{s_2}e^{2z}\right)^{-\frac{s_1+s_2}{2}}$$

$$(-\infty<z<+\infty),$$

care este tocmai *legea de probabilitate a variabilei* z, care se prezintă în numeroase aplicaţii.

§ 13. Distribuția Beta

Fiind dată variabila

$$u=\frac{x_1^2+x_2^2+\ldots+x_{s_1}^2}{x_1^2+x_2^2+\ldots+x_{s_1}^2+y_1^2+y_2^2+\ldots+y_{s_2}^2}, \qquad \text{(XI.13,1)}$$

unde $x_i\,(i=1,\ 2,\ldots,s_1)$ și $y_j\,(j=1,\ 2,\ldots,s_2)$ sînt variabile independente normale $N(0,\ \sigma)$, ne propunem să-i determinăm legea de probabilitate.

1. Vom presupune ca mai înainte că, în spațiul euclidian cu s_1+s_2 dimenisiuni, proiectăm punctul $M(x_1,\ x_2,\ldots,x_{s_1},\ y_1,\ldots,y_{s_2})$ pe subspațiile liniare

$$L_1\,(x_1,\ x_2,\ldots,x_{s_1},\ 0,\ldots,0)$$

și

$$L_2(0,\ 0,\ldots,0,\ y_1,\ldots,y_{s_2}).$$

Considerăm în fiecare dintre spațiile liniare L_1 și L_2 o transformare de variabile în coordonate polare

$$\begin{aligned}
x_1&=\rho\cos\theta\cos\varphi_1\ldots\cos\varphi_{s_1-2}\cdot\cos\varphi_{s_1-1},\\
x_2&=\rho\cos\theta\cdot\cos\varphi_1\ldots\cos\varphi_{s_1-2}\cdot\sin\varphi_{s_1-1},\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
x_{s_1}&=\rho\cdot\cos\theta\cdot\sin\varphi_1\,; \qquad \text{(XI.13,2)}\\
y_1&=\rho\cdot\sin\theta\cdot\cos\psi_1\ldots\cos\psi_{s_2-2}\cdot\cos\psi_{s_2-1},\\
y_2&=\rho\cdot\sin\theta\cdot\cos\psi_1\ldots\cos\psi_{s_2-2}\cdot\sin\psi_{s_2-1},\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
y_{s_2}&=\rho\sin\theta\cdot\sin\psi_1.
\end{aligned}$$

Prin efectul acestor transformări, vor rezulta relațiile

$$x_1^2+x_2^2+\ldots+x_{s_1}^2=\rho^2\cos^2\theta$$

și

$$y_1^2+y_2^2+\ldots+y_{s_2}^2=\rho^2\sin^2\theta,$$

din care deducem expresia

$$u=\cos^2\theta, \qquad \text{(XI.13,3)}$$

a cărei interpretare geometrică este imediată.

Această relaţie ne dă posibilitatea de a găsi legea de probabilitate a variabilei u, deoarece cunoaştem legea de probabilitate elementară a lui θ, care, după cum s-a arătat, este egală cu

$$2\frac{\Gamma\left(\frac{s_1+s_2}{2}\right)}{\Gamma\left(\frac{s_1}{2}\right)\Gamma\left(\frac{s_2}{2}\right)}\cdot\cos^{s_1-1}\theta\cdot\sin^{s_2-1}\theta\cdot d\theta \qquad \left(0\leqslant\theta\leqslant\frac{\pi}{2}\right).$$

Din (XI.13,3) rezultă

$$\cos\theta=u^{\frac{1}{2}},$$

$$\sin\theta=(1-u)^{\frac{1}{2}},$$

$$d\theta=\frac{1}{2}u^{-\frac{1}{2}}(1-u)^{-\frac{1}{2}}du,$$

luînd în membrul al doilea valoarea pozitivă.

De aici, înlocuind fiecare dintre aceste valori, în legea de probabilitate elementară a lui θ de mai sus, găsim

$$\frac{\Gamma\left(\frac{s_1+s_2}{2}\right)}{\Gamma\left(\frac{s_1}{2}\right)\cdot\Gamma\left(\frac{s_2}{2}\right)}\cdot u^{\frac{s_1}{2}-1}\cdot(1-u)^{\frac{s_2}{2}-1}\cdot du \qquad (0\leqslant u\leqslant 1). \qquad \text{(XI.13,4)}$$

Se ştie, însă, că

$$\int_0^1 u^{\frac{s_1}{2}-1}(1-u)^{\frac{s_2}{2}-1}du=B\left(\frac{s_1}{2},\frac{s_2}{2}\right).$$

Atunci, din expresia (XI.13,4) obţinem

$$\frac{1}{B\left(\frac{s_1}{2},\frac{s_2}{2}\right)}\cdot u^{\frac{s_1}{2}-1}\cdot(1-u)^{\frac{s_2}{2}-1}\cdot \qquad \text{(XI.13,5)}$$

Aceasta este *legea de probabilitate a distribuţiei Beta.*

2. La distribuţia Beta se poate ajunge plecînd de la legea de probabilitate a distribuţiei Snedecor cu s_1 şi s_2 grade de libertate.
In adevăr, prin substituţia

$$v = \frac{\frac{s_1}{s_2} u}{1 + \frac{s_1}{s_2} u},$$

legea de probabilitate (XI.12,4) a distribuţiei Snedecor

$$\left(\frac{s_1}{s^2}\right)^{\frac{s_1}{2}} \cdot \frac{\Gamma\left(\frac{s_1+s_2}{2}\right)}{\Gamma\left(\frac{s_1}{2}\right)\cdot\Gamma\left(\frac{s_2}{2}\right)} \cdot u^{\frac{s_1}{2}-1} \cdot \left(1+\frac{s_1}{s_2}u\right)^{-\frac{s_1+s_2}{2}},$$

se va scrie

$$\frac{\Gamma\left(\frac{s_1+s_2}{2}\right)}{\Gamma\left(\frac{s_1}{2}\right)\cdot\Gamma\left(\frac{s_2}{2}\right)} v^{\frac{s_1}{2}-1}(1-v)^{\frac{s_2}{2}-1} = \frac{1}{B\left(\frac{s_1}{2}, \frac{s_2}{2}\right)} \cdot v^{\frac{s_1}{2}-1} \cdot (1-v)^{\frac{s_2}{2}-1},$$

adică aceeaşi lege de probabilitate (XI.13,5) pe care am găsit-o şi pe cale directă.
Momentele distribuţiei Beta rezultă imediat

$$M_r = \frac{1}{B\left(\frac{s_1}{2}, \frac{s_2}{2}\right)} \int_0^1 u^r u^{\frac{s_1}{2}-1}(1-u)^{\frac{s_2}{2}-1} du =$$

$$= \frac{1}{B\left(\frac{s_1}{2}, \frac{s_2}{2}\right)} \int_0^1 u^{r+\frac{s_1}{2}-1}(1-u)^{\frac{s_2}{2}-1} du = \frac{B\left(r+\frac{s_1}{2}, \frac{s_2}{2}\right)}{B\left(\frac{s_1}{2}, \frac{s_2}{2}\right)}.$$

§ 14. Distribuţia rectangulară

Ne propunem a determina o funcţie de frecvenţă $f(x)$, constantă pentru mulţimea valorilor variabilei x care aparţin unui interval dat $[a, b]$.

Deci, trebuie să găsim valoarea funcției $f(x)$, astfel ca

$$\begin{aligned} f(x)&=0, \qquad x<a,\\ f(x)&=k, \qquad a\leqslant x\leqslant b,\\ f(x)&=0, \qquad x>b. \end{aligned} \tag{XI.14,1}$$

Constanta k se va determina din condiția următoare

$$k\int_a^b dx=1,$$

de unde deducem

$$k=\frac{1}{b-a}.$$

Prin urmare, funcția $f(x)$ definită de egalitățile (XI.14,1) va fi

$$f(x)=\frac{1}{b-a}, \tag{XI.14,2}$$

care este *funcția de frecvență a distribuției rectangulare.*
In particular, dacă $a=0$ și $b=1$, funcția de frecvență de mai sus devine

$$f(x)=1.$$

Graficul funcției de frecvență $f(x)$, în baza relației (XI.14,2) este în intervalul său de existență $[a, b]$ o dreaptă paralelă cu axa x-lor.
Funcția de repartiție este

$$F(x)=P(\xi<x)=\frac{1}{b-a}\int_a^x d\xi=\frac{x-a}{b-a}=kx+k',$$

unde am notat $k=\frac{1}{b-a}$ ca mai sus și $k'=-\frac{a}{b-a}$.
Momentele distribuției rectangulare se calculează imediat.
In adevăr, avem

$$M_r=\frac{1}{b-a}\int_a^b x^r\,dx=\frac{1}{r+1}\cdot\frac{b^{r+1}-a^{r+1}}{b-a}.$$

Dînd lui r valorile 1, 2, ..., obţinem primele momente ale distribuţiei rectangulare

$$M_1 = \frac{1}{2}(a+b),$$

$$M_2 = \frac{1}{3}(a^2+ab+b^2),$$

$$M_3 = \frac{1}{4}(a+b)(a^2+b^2),$$

. . . . . .

$$m_2 = \sigma^2 = M_2 - M_1^2 = \frac{1}{12}(b-a)^2.$$

Funcţia caracteristică a acestei distribuţii, conform definiţiei sale, este

$$\varphi(t) = M(e^{itx}) = \frac{1}{b-a}\int_a^b e^{itx}\,dx = \frac{1}{it}\cdot\frac{e^{itb}-e^{ita}}{b-a}.$$

Funcţia de frecvenţă (XI.14,2) a distribuţiei rectangulare este cea mai simplă dintre toate funcţiile de frecvenţă studiate pînă acum; ea prezintă interes pentru unèle aplicaţii ale statisticii matematice.

Iată una dintre aceste aplicaţii referitoare la măsura împrăştierii unei distribuţii.

Din relaţiile

$$M_1 = m = \frac{b+a}{2}, \quad \sigma = \frac{b-a}{2\sqrt{3}},$$

deducem

$$b = m+\sigma\sqrt{3}, \quad a = m-\sigma\sqrt{3},$$

$$f(x) = \frac{1}{2\sigma\sqrt{3}}.$$

Să considerăm o distribuţie oarecare avînd media m şi dispersia σ Din cele de mai sus se vede că această distribuţie are aceeaşi medie şi dispersie ca distribuţia rectangulară cu funcţia de frecvenţă

$$f(x) = \frac{1}{2\sigma\sqrt{3}},$$

Scanned with OKEN Scanner

definită în intervalul $(m-\sigma\sqrt{3},\ m+\sigma\sqrt{3})$. Prin urmare, distribuţia rectangulară ne poate servi pentru măsurarea împrăştierii distribuţiei considerate. Această observaţie nu are vreo valoare pentru o dimensiune, însă poate fi utilă pentru construirea unui indicativ al împrăştierii masei pentru distribuţiile cu mai multe dimensiuni. Va trebui pentru aceasta să definim mai întîi distribuţia rectangulară cu mai multe dimensiuni.

Intr-un spaţiu euclidian cu k dimensiuni V_k să considerăm o regiune finită R din spaţiu şi să determinăm funcţia de frecvenţă $f(x)$, prin x notînd un punct din V_k, astfel ca

$$f(x)=\alpha \quad \text{pentru} \quad x\in R,$$

$$f(x)=0 \quad \text{pentru} \quad x\in\overline{R},$$

prin $\overline{R}$ înţelegînd complementara lui R.

Din condiţia

$$\alpha \iint\limits_{R} \dots \int dx_1\, dx_2 \dots dx_n = 1,$$

determinăm pe α.

Aceasta este o distribuţie rectangulară în spaţiul V_k.

Să considerăm acum o distribuţie oarecare cu media $m\ (m_1, m_2, \dots, m_k)$ şi momentele de ordinul al doilea $m_{ij}\ (i, j=1, 2, \dots, k)$. Nu vom produce nici o particularizare dacă $m_1=m_2=\dots=m_n=0$.

Dacă $A=\|a_{ij}\|\ (i, j=1, \dots, k)$ este o matrice pătrată, de rangul k şi forma pătratică $x'Ax$ unde x este vectorul

$$x=\begin{Vmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{Vmatrix},$$

este definită pozitivă, ecuaţia

$$x'Ax=1$$

defineşte în V_k un elipsoid E cu centrul în origine şi cu volumul V. Dacă în interiorul elipsoidului luăm o distribuţie egală cu $\frac{1}{V}$ şi în exteriorul lui egală cu zero, definim o distribuţie rectangulară în V_k. Se ştie că

$$V=\frac{\pi^{\frac{k}{2}}}{\sqrt{|A|}}\cdot\frac{1}{\Gamma\left(\frac{k}{2}+1\right)}.$$

Ne propunem să determinăm pe A în aşa fel încît să avem

$$\frac{1}{V}\int\int\ldots\int_V x_i x_j dx_1 dx_2\ldots dx_k=m_{ij}\qquad(i,j=1,2,\ldots,k).$$

Cu alte cuvinte, alegem elipsoidul E astfel încît distribuţia rectangulară definită cu ajutorul lui, în modul mai sus indicat, să aibă momentele de primele două ordine egale cu momentele distribuţiei considerate.

Dacă u este vectorul

$$u=\left\|\begin{matrix}u_1\\u_2\\\cdot\\\cdot\\\cdot\\u_k\end{matrix}\right\|$$

din relaţia precedentă, rezultă

$$\frac{1}{V}\int\int\ldots\int_E(u'x)^2dx_1dx_2\ldots dx_k=u'Mu,$$

prin M însemnînd matricea covariantă $M=\|m_{ij}\|_{i,j=1,2,\ldots,k}$.

Să efectuăm o transformare ortogonală

$$x=Hy,$$

astfel ca forma pătratică $x'Ax$ să se transforme în

$$y'H'AHy=\sum_{i=1}^{k}\lambda_i y_i^2,$$

λ_i fiind constante pozitive, diferite de zero.

Prin această transformare, vectorul u se transformă în vectorul v pe baza relației $v=H'u$ care, ținînd seama că $H'=H^{-1}$, matricea H fiind ortogonală, ne dă $u=Hv$.

Forma pătratică $u'Mu$ devine $v'H'MHv$.

Avem succesiv

$$v'H'MHv=\frac{1}{V}\int\int\dots\int_{x'Ax<1}(u'x)^2dx_1\dots dx_k=$$

$$=\frac{1}{V}\int\int\dots\int_{\sum_{i=1}^{k}\lambda_i y_i^2<1}(v'y)^2\,dy_1dy_2\dots dy_k=$$

$$=\frac{1}{V}\sum_{i=1}^{k}v_i^2\int\int\dots\int_{\sum_{i=1}^{k}\lambda_i y_i^2<1}y_i^2\,dy_1dy_2\dots dy_k,$$

deoarece, din cauza simetriei, integralele

$$\int\int\dots\int y_iy_j\,dy_1dy_2\dots dy_k$$

sînt nule.

Pe de altă parte, dacă efectuăm substituția

$$y_i=\frac{z_i}{\sqrt{\lambda_i}}\qquad(i=1,2,\dots,k),$$

rezultă

$$\frac{v_i^2}{V}\int\int\dots\int_{\sum_{i=1}^{k}\lambda_i y_i^2<1}y_i^2dy_1dy_2\dots dy_k=$$

$$=\frac{v_i^2}{V\lambda_i\sqrt{\lambda_1\lambda_2\dots\lambda_k}}\cdot\frac{\pi^{\frac{k}{2}}}{(k+2)\,\Gamma\left(\frac{k}{2}+1\right)}=\frac{v_i^2}{(k+2)\,\lambda_i},$$

ținînd seama de valoarea lui V dată mai sus de relația

$$\lambda_1\lambda_2\dots\lambda_k=|A|,$$

precum şi de valoarea integralei

$$\int\int\ldots\int_{\sum_{i=1}^{k} z_i^2<1} z_i^2 dz_1 dz_2 \ldots dz_k.$$

Avem

$$v'H'AHv = \sum_{i=1}^{k} \lambda_i v_i^2,$$

deci

$$v'(H'AH)^{-1}v = \sum_{i=1}^{k} \frac{v_i^2}{\lambda_i},$$

$$v'H'MHv = \frac{1}{k+2} v'(H'AH)^{-1}v,$$

$$H'MH = \frac{1}{k+2}(H'AH)^{-1},$$

$$H'AH \cdot H'MH = \frac{1}{k+2} E.$$

Cum

$$HH' = HH^{-1} = E,$$

rezultă mai departe

$$H^{-1}AMH = \frac{E}{k+2}, \quad A = \frac{M^{-1}}{k+2}.$$

Prin urmare, elipsoidul căutat are ecuaţia

$$\frac{1}{k+2} x'M^{-1}x = 1,$$

M fiind matricea covariantă.

El poartă numele de *elipsoid de concentrare* şi constituie un indicator pentru măsurarea împrăştierii în jurul mediei la distribuţiile cu acelaşi număr de dimensiuni.

Dacă pentru două distribuţii cu acelaşi număr de dimensiuni şi aceeaşi medie, elipsoidele de concentrare sînt astfel aşezate, ca elipsoidul primei distribuţii să fie cuprins în întregime în elipsoidul distribuţiei a doua, înseamnă că prima distribuţie are o concentrare mai mare de mase în jurul mediei decît a doua.

Scanned with OKEN Scanner

In particular, pentru $k=2$, nu avem o elipsă de concentrare, a cărei ecuaţie se determină din ecuaţia generală a elipsoidului de inerţie. Notînd, ca de obicei

$$m_{20}=\sigma_1^2, \quad m_{11}=\rho\sigma_1\sigma_2, \quad m_{02}=\sigma_2^2,$$

matricea covariantă este

$$M=\begin{Vmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{Vmatrix},$$

deci

$$M^{-1}=\begin{Vmatrix} \frac{1}{(1-\rho^2)\sigma_1^2} & -\frac{\rho}{(1-\rho^2)\sigma_1\sigma_2} \\ -\frac{\rho}{(1-\rho^2)\sigma_1\sigma_2} & \frac{1}{(1-\rho^2)\sigma_2^2} \end{Vmatrix}.$$

Ecuaţia elipsei de concentrare este

$$\frac{1}{4}x'M^{-1}x=1$$

sau

$$\frac{x_1^2}{4(1-\rho^2)\sigma_1^2}-\frac{\rho x_1x_2}{2(1-\rho^2)\sigma_1\sigma_2}+\frac{x_2^2}{4(1-\rho^2)\sigma_2^2}=1.$$

§ 15. Distribuţia Cauchy

Prin definiţie, o distribuţie a cărei funcţie de frecvenţă este egală cu

$$\frac{1}{\pi}\cdot\frac{1}{1+x^2} \qquad (-\infty<x<+\infty) \qquad \text{(XI.15,1)}$$

se numeşte *distribuţie Cauchy*.

Funcţia dată de expresia (XI.15,1) este în adevăr o funcţie de frecvenţă, deoarece ea este pozitivă pentru orice valoare a lui x cuprinsă între $-\infty$ şi $+\infty$, iar

$$\frac{1}{\pi}\int_{-\infty}^{+\infty}\frac{dx}{1+x^2}=\frac{1}{\pi}(\operatorname{arctg}x)_{-\infty}^{+\infty}=1.$$

Este de observat că distribuţia Cauchy nu are momente de nici un ordin, inclusiv momentul de ordinul întîi.
Funcţia sa caracteristică există şi este

$$\varphi(t)=e^{-|t|},$$

după cum se poate constata uşor.
Potrivit unei formule cunoscute

$$f(x)=\frac{1}{2\pi}\int_{-\infty}^{+\infty}\varphi(t)\,e^{-itx}dt.$$

Inlocuind în această relaţie pe $\varphi(t)$ cu $e^{-|t|}$, obţinem

$$f(x)=\frac{1}{2\pi}\int_{-\infty}^{+\infty}e^{-|t|}\,e^{-itx}\,dt=\frac{1}{2\pi}\left(\int_{-\infty}^{0}e^{t-itx}\,dt+e^{-t-itx}\,dt\right)=$$

$$=\frac{1}{2\pi}\left(\int_{0}^{\infty}e^{-t-itx}\,dt-\int_{\infty}^{0}e^{-t+itx}\,dt\right)=\frac{1}{2\pi}\int_{0}^{\infty}e^{-t}\,(e^{itx}+e^{-itx})\,dt=$$

$$=\frac{1}{\pi}\int_{0}^{\infty}e^{-t}\cos tx\,dt.$$

Integrînd succesiv prin părţi de două ori ultima integrală, rezultă

$$\frac{1}{\pi}\int_{0}^{\infty}e^{-t}\cos tx\,dt=\frac{1}{\pi}-\frac{x^2}{\pi}\int_{0}^{\infty}e^{-t}\cos tx\,dt.$$

sau

$$f(x)=\frac{1}{\pi}\cdot\frac{1}{1+x^2}.$$

Am găsit, astfel, funcţia de frecvenţă (XI.15,3) a distribuţiei Cauchy.

§ 16. Distribuţia Wishart

Distribuţia normală cu o singură dimensiune a fost, după cum am văzut, punctul de plecare pentru găsirea altor distribuţii clasice din statistica matematică. Astfel, considerînd n variabile independente x_i $(i=1,2,\ldots,n)$, fiecare urmînd o distribuţie normală $N(0,\sigma)$, am arătat că suma pătratelor lor urmează legea χ^2 a lui Helmert-Pearson. Generalizarea firească a acestei din urmă distribuţii se va face considerînd n variabile independente x_i' $(i=1,2,\ldots,n)$, fiecare urmînd

aceeaşi distribuţie normală cu mai multe dimensiuni şi determinînd legea mixtă de probabilitate a sumelor pătratelor componentelor variabilelor $x_i\,(i=1,2,\ldots,n)$ şi a produselor componentelor acestor variabile.

Mai precis, presupunem că variabilele $x_i\,(i=1,2,\ldots,n)$, independente între ele, admit fiecare distribuţia normală

$$\frac{|A|^{\frac{1}{2}}}{(2\pi)^{\frac{s}{2}}}\,e^{-\frac{1}{2}x'Ax}, \qquad \text{(XI.16,1)}$$

prin $A=\|a_{ij}\|$ $(i=1,2,\ldots,s)$ notînd matricea pătrată cu elementele a_{ij}. Ştim că această matrice este definită pozitivă.

Notăm prin $x_{1i}, x_{2i},\ldots,x_{si}$ componentele variabilei x_i $(i=1,2,\ldots,n)$. Deoarece i variază de la 1 la n, avem de considerat ns variabile aleatorii

$$\left\|\begin{array}{cccc} x_{11} & x_{21} & \ldots & x_{s1} \\ x_{12} & x_{22} & \ldots & x_{s2} \\ \cdot & \cdot & \cdot & \cdot \\ x_{1n} & x_{2n} & \ldots & x_{sn} \end{array}\right\|. \qquad \text{(XI.16,2)}$$

Variabilele din fiecare linie admit legea de probabilitate definită de (XI.16,1). Să notăm prin X_k vectorul

$$X_k=\left\|\begin{array}{c} x_{k1} \\ x_{k2} \\ \cdot \\ \cdot \\ \cdot \\ x_{kn} \end{array}\right\|$$

obţinut prin considerarea coloanelor din matricea precedentă.

Fie

$$y_{ij}=X_i'\cdot X_j \qquad (i,j=1,2,\ldots,s)$$

sau

$$y_{ii}=\sum_{h=1}^{n}x_{ih}^2 \qquad (i=1,2,\ldots,s),$$

$$y_{ij}=\sum_{h=1}^{n}x_{ih}x_{jh} \qquad (i,j=1,2,\ldots,s;\ i\neq j).$$

In total avem s variabile y_{ii} şi $\frac{s(s-1)}{2}$ variabile y_{ij}. Deci, în total, $\frac{s(s+1)}{2}$ variabile, a căror lege mixtă de probabilitate ne propunem s-o aflăm.

Legea mixtă de probabilitate a elementelor matricei (XI.16,2) este

$$\frac{|A|^{\frac{n}{2}}}{(2\pi)^{\frac{ns}{2}}} \cdot e^{-\frac{1}{2}\sum_{i,j=1}^{s} a_{ij} \sum_{h=1}^{n} x_{ih}x_{jh}} \Pi dx_{ih}$$

$$(i=1,2,\ldots,s), \quad (h=1,2,\ldots,n).$$

Să notăm prin

$$\varphi(u)=M\left(e^{i\sum_{i,j=1}^{s} \varepsilon_{ij}u_{ij}y_{ij}}\right)$$

funcţia caracteristică a variabilei y cu

$$\varepsilon_{ij}=1, \qquad \text{pentru} \quad i=j,$$

$$\varepsilon_{ij}=\frac{1}{2}, \qquad \text{pentru} \quad i\neq j.$$

Rezultă că

$$\varphi(u)=\frac{|A|^{\frac{n}{2}}}{(2\pi)^{\frac{ns}{2}}} \int\int\ldots\int_{R_n} e^{-\frac{1}{2}\sum_{i,j=1}^{s}(a_{ij}-2i\varepsilon_{ij}u_{ij})y_{ij}} \Pi dx_{ij}$$

$$(i=1,2,\ldots,s; \quad j=1,2,\ldots,n).$$

Valoarea acestei integrale este dată de formula (X.16,4). Obţinem astfel

$$\varphi(u)=\frac{|A|^{\frac{n}{2}}}{|B|^{\frac{n}{2}}}, \qquad \text{(XI.16,4)}$$

unde B este matricea

$$B=\| a_{ij}-2i\varepsilon_{ij}u_{ij} \|.$$

Distribuţia corespunzătoare acestei funcţii caracteristice rezultă în mod simplu şi astfel, din relaţia

$$\iint\dots\int_{|X|\geqq 0} |X|^m e^{-\frac{1}{2}\sum a_{ij}x_{ij}} \prod_{(i,j=1,2,\dots,s)} dx_{ij} = \frac{\pi^{\frac{s(s-1)}{4}} 2^{s\left(m+\frac{s+1}{2}\right)} \Gamma\left(m+\frac{s+1}{2}\right)\dots\Gamma(m+1)}{|A|^{m+\frac{s+1}{2}}}, \qquad \text{(XI.16,5)}$$

unde

$$m \geqq 0,\quad X=\left\| \begin{array}{cccc} x_{11} & x_{12} & \dots & x_{1s} \\ x_{21} & x_{22} & \dots & x_{2s} \\ \cdot & \cdot & \cdot & \cdot \\ x_{s1} & x_{s2} & \dots & x_{ss} \end{array} \right\|, \qquad x_{ij}=x_{ji}$$

şi $A=\| a_{ij} \|$ este o matrice definită pozitivă, rezultă că putem defini o distribuţie $F(m, s)$ în spaţiul $R_{\frac{s(s+1)}{2}}$, a cărei funcţie de frecvenţă este egală cu

$$\frac{|A|^{m+\frac{s+1}{2}}}{\pi^{\frac{s(s-1)}{4}} 2^{s\left(m+\frac{s+1}{2}\right)} \Gamma\left(m+\frac{s+1}{2}\right)\dots\Gamma(m+1)} X^m e^{-\frac{1}{2}\sum a_{ij}x_{ij}} \quad \text{în regiunea } X \geqq 0$$

$$0 \quad \text{în regiunea } X < 0.$$

Funcţia caracteristică a acestei distribuţii este

$$\frac{|A|^{m+\frac{s+1}{2}}}{\pi^{\frac{s(s-1)}{4}} 2^{s\left(m+\frac{s+1}{2}\right)} \Gamma\left(m+\frac{s+1}{2}\right)\dots\Gamma(m+1)} \times$$

$$\times \iint\dots\int |X|^m e^{-\frac{1}{2}\sum_{i,j=1}^{s} (a_{ij}-2it_{ij}u_{ij})x_{ij}} \prod_{(i,j=1,2,\dots,s)} dx_{ij} = \frac{|A|^{m+\frac{s+1}{2}}}{|B|^{m+\frac{s+1}{2}}},$$

unde $B=\| a_{ij}-2it_{ij}u_{ij} \|$.

Pentru a găsi funcţia de frecvenţă a distribuţiei căutate este deci suficient ca în distribuţia $F(m, s)$ să facem

$$m+\frac{s+1}{2}=\frac{n}{2}.$$

Rezultă că distribuţia variabilelor y_{ii} şi y_{ij} are funcţia de frecvenţă

$$\frac{|A|^{\frac{n}{2}}}{\pi^{\frac{s(s-1)}{4}}\cdot 2^{\frac{ns}{2}}\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{n-1}{2}\right)\ldots\Gamma\left(\frac{n-s+1}{2}\right)}|Y|^{\frac{n-s+1}{2}}e^{-\frac{1}{2}\sum a_{ij}x_{ij}}\prod dy_{ij},$$

$$(i,\ j=1,\ 2,\ldots,\ s)$$

prin Y notînd matricea $\|y_{ij}\|$ $(i,\ j=1,\ 2,\ldots,\ s)$.

Ea poartă numele de *distribuţia lui Wishart*, după numele statisticianului care a descoperit-o (1928). Numărul gradelor de libertate ale acestei distribuţii este n.

Pentru $s=1$, obţinem distribuţia χ^2.

Distribuţia Wishart este stabilă.

Să considerăm astfel în $R_{\frac{s(s+1)}{2}}$ două variabile aleatorii independente ξ şi η, care urmează fiecare o lege Wishart cu aceeaşi matrice A. Funcţiile caracteristice ale acestor variabile sînt

$$\varphi_1(u)=\frac{|A|^{\frac{n_1}{2}}}{|B|^{\frac{n_1}{2}}},\quad \varphi_2(u)=\frac{|A|^{\frac{n_2}{2}}}{|B|^{\frac{n_2}{2}}}.$$

Funcţia caracteristică a variabilei $\xi+\eta$ este

$$\varphi_1(u)\cdot\varphi_2(u)=\frac{|A|^{\frac{n_1+n_2}{2}}}{|B|^{\frac{n_1+n_2}{2}}},$$

adică funcţia caracteristică a unei distribuţii Wishart.

Proprietatea este deci demonstrată.

Scanned with OKEN Scanner

§ 17. Distribuția lui Hotelling

Să considerăm acum $(n+i)$ variabile independente x_i $(i=1, 2, \dots, n+1)$, fiecare variabilă x_i având componentele $x_{1i}, x_{2i}, \dots, x_{si}$.
Variabilele x_i sînt distribuite normal.
Notăm

$$y_{ij}=\sum_{h=1}^{n} x_{ih}x_{jh} \qquad (i, j=1, 2, \dots, s),$$

$$z_{ij}=\sum_{h=1}^{n+1} x_{ih}x_{jh} \qquad (i, j=1, 2, \dots, s),$$

$$Y=\| y_{ij} \|,$$

$$Z=\| z_{ij} \|.$$

Ne propunem să determinăm distribuția variabilei

$$v=\frac{|Y|}{|Z|}.$$

Pentru a simplifica calculele, vom însemna

$$A_p=\pi^{\frac{s(s-1)}{4}} \cdot 2^{ps}\,\Gamma(p)\,\Gamma\left(p-\frac{1}{2}\right)\dots\Gamma\left(p-\frac{s-1}{2}\right). \qquad \text{(XI.17,1)}$$

Cu această notație, formula (XI.16,5) devine

$$\int\int\dots\int |X|^m e^{-\frac{1}{2}\sum a_{ij}x_{ij}} \prod_{(i, j=1, 2, \dots, s)} dx_{ij}=\frac{A_{m+\frac{s+1}{2}}}{|A|^{m+\frac{s+1}{2}}}, \qquad \text{(XI.17,2)}$$

iar distribuția lui Wishart cu n grade de libertate devine

$$\frac{|A|^{\frac{n}{2}}}{A_{\frac{n}{2}}} |Y|^{\frac{n-s-1}{2}} e^{-\frac{1}{2}\sum a_{ij}y_{ij}} \prod_{(i, j=1, 2, \dots, s)} dy_{ij}. \qquad \text{(XI.17,3)}$$

Aceasta este, după cum am văzut în paragraful precedent, distribuția variabilelor y_{ij} $(i, j=1, 2, \dots, s)$ în spațiul $R_{\frac{s(s+1)}{2}}$.

In acelaşi spaţiu, distribuţia variabilelor z_{ij} se va obţine din formula (XI.17,3), punînd $n+1$ în loc de n.
Avem

$$M(|Y|^{\lambda})=\frac{|A|^{\frac{n}{2}}}{A_{\frac{n}{2}}}\int\int\cdots\int|Y|^{\lambda+\frac{n-s-1}{2}}e^{-\sum a_{ij}y_{ij}}\prod dy_{ij}.$$

Ţinînd seamă de (XI.17,2), rezultă mai departe

$$M(|Y|^{\lambda})=\frac{A_{\lambda+\frac{n}{2}}}{A_{\frac{n}{2}}}\cdot\frac{1}{|A|^{\lambda}}, \qquad \text{(XI.17,4)}$$

cu condiţia ca

$$\min[n;n+2\lambda]>s.$$

La fel rezultă

$$M(|Z|^{\mu})=\frac{A_{-\mu+\frac{n+1}{2}}}{A_{\frac{n+1}{2}}}\cdot\frac{1}{|A|^{\mu}}, \qquad \text{(XI.17,5)}$$

cu condiţia ca

$$\min[n+1;n+1+2\mu]>s.$$

Putem însă scrie

$$Z_{ij}=y_{ij}+x_{i,n+1}x_{j,n+1} \qquad (i,j=1,2,\ldots,s).$$

Variabilele y_{ij} şi $x_{i,n+i}$ sînt independente. Funcţia de frecvenţă a distribuţiei variabilelor y_{ij} este dată de (XI.17,3), iar variabila x_{n+i} urmează o distribuţie normală cu s dimensiuni.
Avem deci

$$M[|Z|^{\mu}]=\frac{|A|^{\frac{n+1}{2}}}{(2\pi)^{\frac{s}{2}}A_{\frac{n}{2}}}\int\int\cdots$$

$$\cdots\int_{|Y|>0}|Z|^{\mu}|Y|^{\frac{n-s-1}{2}}e^{-\frac{1}{2}\sum a_{ij}y_{ij}}e^{-\frac{1}{2}\sum a_{ij}x_{i,n+i}x_{j,n+i}}\times$$

$$\times\prod_{i,j=1}^{s}dy_{ij}\prod_{i=1}^{s}dx_{i,n+i}.$$

Comparînd cu (XI.17,5), obținem astfel

$$\iint\cdots\int_{|Y|>0} |Z|^{\mu}|Y|^{\frac{n-s-1}{2}} e^{-\frac{1}{2}\sum a_{ij}y_{ij}} e^{-\frac{1}{2}\sum a_{ij}r_{i,n+t}x_{j,n+t}} \times$$

$$\times \prod dy_{ij} \prod dx_{i,n+i} = (2\pi)^{\frac{s}{2}} \frac{A_{\frac{n}{2}} A_{\mu+\frac{n+1}{2}}}{A_{\frac{n+1}{2}}} |A|^{-\frac{n+1}{2}-\mu}, \qquad \text{(XI.17,6)}$$

unde

$$\min(n,\ n+2\mu+1) > s.$$

Insă

$$M[|Y|^{\lambda}|Z|^{\mu}] = \frac{|A|^{\frac{n+1}{2}}}{(2\pi)^{\frac{s}{2}} A_{\frac{n}{2}}} \iint\cdots\int |Z|^{\mu}|Y|^{\lambda+\frac{n-s-1}{2}} \times$$

$$\times e^{-\frac{1}{2}\sum a_{ij}y_{ij}-\frac{1}{2}\sum a_{ij}x_{i,n+t}x_{j,n+t}} \prod dy_{ij} \prod dx_{i,n+i}. \qquad \text{(XI.17,7)}$$

Sau, ținînd seamă de (XI.17,6), în care în locul lui n punem $n+2\lambda$

$$M[|Y|^{\lambda}|Z|^{\mu}] = \frac{A_{\frac{n}{2}+\lambda}}{A_{\frac{n}{2}}} \cdot \frac{A_{\lambda+\mu+\frac{n+1}{2}}}{A_{\lambda+\frac{n+1}{2}}} \cdot \frac{1}{|A|^{\lambda+\mu}}. \qquad \text{(XI.17,8)}$$

Egalitatea are loc pentru $\lambda \geqq 0$ și $n+1+2\lambda+2\mu \geqq s$.
Putem deci face $\mu = -\lambda$.
Obținem astfel

$$M\left(\frac{|Y|^{\lambda}}{|Z|^{\lambda}}\right) = \frac{A_{\frac{n}{2}+\lambda}}{A_{\frac{n}{2}}} \cdot \frac{A_{\frac{n+1}{2}}}{A_{\frac{n+1}{2}+\lambda}}.$$

Ținînd seama de (XI.17,1), relația precedentă devine

$$M\left(\frac{|Y|^{\lambda}}{|Z|^{\lambda}}\right) = \frac{\Gamma\left(\frac{n-s+1}{2}+\lambda\right)}{\Gamma\left(\frac{n+1}{2}+\lambda\right)} \cdot \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n-s+1}{2}\right)} = \frac{B\left(\lambda+\frac{n-s+1}{2}, \frac{s}{2}\right)}{B\left(\frac{n-s+1}{2}, \frac{s}{2}\right)}.$$

Momentul de ordinul λ al variabilei

$$v=\frac{|Y|}{|Z|}$$

coincide prin urmare cu momentul de ordinul λ al distribuţiei Beta (cap. XI, § 13). Condiţiile de unicitate din problema momentelor fiind îndeplinite, rezultă că distribuţia lui v este distribuţia Beta a cărei funcţie de frecvenţă este

$$\frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{s}{2}\right)\Gamma\left(\frac{n-s+1}{2}\right)}\,v^{\frac{n-s+1}{2}}(1-v)^{\frac{s}{2}-1}. \qquad \text{(XI.17,9)}$$

$$0\leqq v\leqq 1.$$

Dacă punem

$$v=\frac{1}{1+\frac{T^2}{n}} \qquad (T\geqq 0), \qquad \text{(XI.17,10)}$$

distribuţia lui T se determină uşor din (XI.17,8).
Fie $G(x)$ funcţia de repartiţie a lui T. Avem

$$G(x)=P[T\leqq x]=P\left[v\geqq\frac{n}{x^2+n}\right]=1-F\left(\frac{n}{x^2+n}\right),$$

prin $F(x)$ însemnînd funcţia de repartiţie a lui v.
Prin derivare rezultă

$$g(T)=\frac{2nT}{(T^2+n)^2}\,f\left(\frac{n}{T^2+n}\right),$$

prin $f(y)$ însemnînd funcţia de frecvenţă (XI.17,8).
Efectuînd calculele, obţinem

$$g(T)=\frac{2\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n}\cdot\Gamma\left(\frac{s}{2}\right)\Gamma\left(\frac{n-s+1}{2}\right)}\cdot\frac{\left(\frac{T^2}{n}\right)^{\frac{s-1}{2}}}{\left(1+\frac{T^2}{n}\right)^{\frac{n+1}{2}}}\cdot \qquad (T>0).$$

Aceasta este *distribuţia lui Hotelling cu n grade de libertate.*
Ea constituie o generalizare a distribuţiei lui Student în spaţiul $R_{\frac{s(s+1)}{2}}$.

Scanned with OKEN Scanner

§ 18. Teorema lui Rao

K. S. Rao a demonstrat că variabilele $|Z|$ şi $\frac{|Y|}{|Z|}$ sînt independente.

Pentru aceasta, este suficient să arătăm că, dacă α şi β sînt două numere întregi oarecare, avem

$$M\left(|Z|^{\alpha}\frac{|Y|^{\beta}}{|Z|^{\beta}}\right)=M(|Z|^{\alpha})\,M\left(\frac{|Y|^{\beta}}{|Z|^{\beta}}\right). \qquad \text{(XI.18,1)}$$

In baza relaţiilor (XI.17,4), (XI.17,5), (XI.17,7) obţinem

$$M\left(|Z|^{\alpha}\frac{|Y|^{\beta}}{|Z|^{\beta}}\right)=M(|Y|^{\beta}|Z|^{\alpha-\beta})=\frac{A_{\frac{n}{2}+\beta}}{A_{\frac{n}{2}}}\cdot\frac{A_{\frac{n+1}{2}+\alpha}}{A_{\frac{n+1}{2}+\beta}}\cdot\frac{1}{|A|^{\alpha}},$$

$$M(|Z|^{\alpha})\,\frac{A_{\frac{n+1}{2}+\alpha}}{A_{\frac{n+1}{2}}}\cdot\frac{1}{|A|^{\alpha}},$$

$$M\left(\frac{|Y|^{\beta}}{|Z|^{\beta}}\right)=\frac{A_{\frac{n}{2}+\beta}}{A_{\frac{n}{2}}}\cdot\frac{A_{\frac{n+1}{2}}}{A_{\frac{n+1}{2}+\beta}}.$$

Inlocuind aceste valori în (XI.18,1) aceasta rezultă imediat.

De aici putem trage o concluzie interesantă referitoare la şirul x_i $(i=1, 2, \ldots, n)$ de variabile normale în R_s.

Să notăm matricea Y din paragraful precedent prin Y_n, pentru a pune în evidenţă că această matrice corespunde la n variabile x_i $(i=1, 2, \ldots, n)$.

Pentru şirul de variabile $x_i\,(i=1, 2, \ldots, p)\,(p\leq n)$ vom avea atunci matricea corespunzătoare Y_p.

Variabilele

$$\frac{|Y_s|}{|Y_{s+1}|},\ \frac{|Y_{s+1}|}{|Y_{s+2}|},\ \ldots,\ \frac{|Y_{n-1}|}{|Y_n|},\ |Y_n| \qquad \text{(XI.18,2)}$$

sînt independente între ele, după cum se poate arăta uşor.

In acelaşi mod cum am arătăt că variabilele $|Z|$ şi $\frac{|Y|}{|Z|}$ sînt independente între ele, adică printr-o simplă verificare, se arată că variabilele $|Y_q|$ şi $\frac{|Y_p|}{|Y_q|}$, $q < p \leq n$, sînt de asemenea independente.

Rezultă atunci că $|Y_{p+1}|$ şi $\frac{|Y_p|}{|Y_{p+1}|}$ sînt independente. Însă

$$\frac{|Y_p|}{|Y_{p+1}|} = \frac{|Y_p|}{|Y_{p+1}|} \cdot \frac{|Y_{p+1}|}{|Y_{p+2}|}.$$

In baza observaţiei precedente, fiecare dintre variabilele $\frac{|Y_p|}{|Y_{p+1}|}$, $\frac{|Y_{p+1}|}{|Y_{p+2}|}$ sînt independente de $|Y_{p+2}|$. Prin urmare şi raportul lor, adică $\frac{|Y_p|}{|Y_{p+1}|}$ este independent de $|Y_{p+2}|$.

Variabila $\frac{|Y_p|}{|Y_{p+1}|}$ fiind independentă de $|Y_{p+1}|$ şi $|Y_{p+2}|$ este independentă de raportul lor

$$\frac{|Y_{p+1}|}{|Y_{p+2}|}.$$

Variabilele din şirul (XI.18,2) sînt deci independente.

Să notăm atunci

$$\frac{|Y_p|}{|Y_{p+1}|} = \frac{1}{1+\frac{T_p^2}{p}} \qquad (p=s,\ s+1,\ \ldots,\ n-1).$$

Variabilele T_p, după cum rezultă din paragraful precedent, urmează fiecare o distribuţie Hotelling respectiv cu p grade de libertate. Mai mult decît atît, din relaţia de mai sus reiese că variabilele T_p sînt independente între ele.

Obţinem astfel următoarea teoremă, datorită lui K. S. Rao.

Scanned with OKEN Scanner

Teoremă. *Fie* $x_1, x_2, \ldots, x_n$, *n variabile normale, independente, din* R_s, *a căror funcție de frecvență este*

$$\frac{|A|^{\frac{1}{2}}}{(2\pi)^{\frac{s}{2}}} e^{-\frac{1}{2}x'Ax},$$

unde $A = \| a_{ij} \|$ *este o matrice pătrată cu s linii și s coloane, pozitiv definită de rangul s.*
Fie $(x_{1k}, x_{2k}, \ldots, x_{sk})$ *componentele lui* x_k $(k=1, 2, \ldots, n)$.
Notăm, ca de obicei, prin x_k *matricea coloană*

$$\left\| \begin{array}{c} x_{1k} \\ x_{2k} \\ \cdot \\ \cdot \\ \cdot \\ x_{sk} \end{array} \right\|$$

și însemnăm prin Y_p *matricea*

$$Y_p = \sum_{\alpha=1}^{p} \| x'_\alpha x_\alpha \| .$$

Variabilele $T_s, T_{s+1}, \ldots, T_{n-1}$, *definite prin egalitățile*

$$\frac{|Y_p|}{|Y_{p+1}|} = \frac{1}{1+\frac{T_p^2}{p}},$$

sînt independente între ele și urmează fiecare o distribuție Hotelling, respectiv cu $s, s+1, \ldots, n-1$ *grade de libertate.*

PARTEA A DOUA

Capitolul XII

Teoria selecției

§ 1. Principiul metodei selecției

O colectivitate statistică sau o populație, după o denumire frecvent întrebuințată în statistică, este, în general, caracterizată, din punct de vedere statistic, prin distribuția ei. De exemplu, piesele lucrate de o anume mașină, pentru a fi acceptate trebuie să îndeplinească anumite condiții. Piesele acceptabile și neacceptabile se distribuie după o lege de distribuție, pe care presupunem că o cunoaștem în condițiile normale de funcționare ale mașinii. Piesele formează o populație statistică, iar legea ne dă distribuția acestei populații. Cunoscînd distribuția, putem determina valoarea medie a numărului de piese acceptabile produse într-un interval de timp dat, împrăștierea lor în jurul valorii medii, care poate fi considerată ca o piesă standard etc. De asemenea, vom putea sezisa momentul cînd mașina încetează de a funcționa în condiții normale. Uneori, din punct de vedere practic, nu ne interesează decît anumite caracteristici ale distribuției, de pildă, primele momente centrate. Problema fundamentală care se pune Statisticii matematice este următoarea: dacă nu știm a priori și nici nu putem determina a priori legea de distribuție a populației, cum putem găsi, pe baza experiențelor ce avem cu populația, această lege sau măcar caracteristicile principale care ne interesează. Vom arăta că, în anumite condiții, lucrul acesta este posibil și reprezintă o metodă importantă de cercetare din statistica matematică, cunoscută sub numele de *metoda selecției* sau a *eșantioanelor*.

In esență, metoda selecției este următoarea: fiind dată o populație formată din N indivizi, în urma a n experiențe independente, bine determinate, obținem rezultatele $x_1, x_2, \ldots, x_n$. Se pune problema determinării caracteristicilor populației (momente, mediană etc.), cu ajutorul variabilelor $x_i (i = 1, 2, \ldots, n)$.

Scanned with OKEN Scanner

Vom numi distribuția populației, *distribuție teoretică,* iar variabilele $x_1, x_2, \ldots, x_n$ *variabile de selecție.* Elementele distribuției teoretice vor fi numite în mod corespunzător, adică media teoretică, mediana teoretică etc. O funcție $f(x_1, x_2, \ldots, x_n)$ poartă numele de *funcție de selecție.* Vom avea, astfel, de considerat numeroase funcții de selecție ca, de exemplu, media de selecție

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n},$$

momentul de selecție de ordinul r

$$\overline{M}_r = \frac{x_1^r + x_2^r + \ldots + x_n^r}{n}$$

etc.

Pentru noi prezintă importanță cazul cînd funcția de frecvență tinde în probabilitate către o limită bine determinată care, potrivit teoremei lui Glivenko, este distribuția teoretică. Rezultă că, din punct de vedere practic, cunoscînd valorile funcției $f(x_1, x_2, \ldots, x_s)$ în urma a n experiențe, care ne-au dat pentru variabilele $x_1, x_2, \ldots, x_n$ niște valori empirice bine determinate, vom putea determina anumite elemente din distribuția teoretică. Și mai precis vom putea proceda în cazul mult mai rar întîlnit în aplicații, cînd cunoaștem distribuția exactă a funcției $f(x_1, x_2, \ldots, x_s)$. Modul cum se pun în aplicație aceste considerații teoretice se va studia în paragraful următor.

Cînd numărul indivizilor din populație este infinit sau poate fi considerat, din punct de vedere practic, ca infinit, fiind foarte mare, singura metodă de cercetare pentru determinarea numerică a diferitelor caracteristici ale distribuției teoretice este metoda selecției. Cînd medicul face o analiză a sîngelui, el aplică metoda selecției. El determină valoarea medie a diferitelor corpuscule din sînge — valoarea medie a acestora fiind în acest caz caracteristica ce ne interesează — cercetînd numai colectivitatea parțială de sînge luat pentru analiză. Fără să aibă nevoie să cunoască principiile statisticii matematice, el nu se îndoiește nici un moment de eficacitatea metodei întrebuințate.

Analize de acest gen, pe baza unei colectivități parțiale, se fac nu numai în medicină, ci în general în tehnică și în științele experimentale. Dar nu întotdeauna se pot trage concluzii exacte atît de ușor ca în exemplul dat mai sus și uneori pentru obținerea unor

rezultate aplicabile în practică, trebuie să se folosească metodele cele mai moderne ale calculului probabilităților.

Metoda selecției se aplică și cînd numărul indivizilor din populație este finit. Să considerăm o producție de N metri de pînză realizată într-un timp dat, într-o fabrică de textile. Ne interesează coeficientul de rezistență al pînzei. Pentru aceasta, aplicăm metoda selecției, luînd la întîmplare diferite bucăți din pînza produsă și determinăm coeficientul de rezistență respectiv.

Aplicînd metodele statisticii matematice, determinăm coeficientul de rezistență al pînzei. Metoda selecției este, în acest caz, singura posibilă din punct de vedere economic, deoarece determinarea directă a coeficientului mediu de rezistență al pînzei ar implica distrugerea întregii producții.

Metoda selecției este folosită și în practica recensămintelor. Să presupunem astfel, că se urmărește determinarea unei caracteristici care ar putea fi obținută pe baza unui recensămînt general, de exemplu, numărul vitelor dintr-o anumită regiune folosite în agricultură. De multe ori, se folosește, pentru realizarea acestui scop un recensămînt parțial, considerîndu-se la întîmplare, numai 5% sau 10% din gospodăriile agricole din regiune. Se ajunge la rezultat pe cale mai rapidă, mai puțin costisitoare și de multe ori și mai exactă, pentru că reducerea materialului statistic de cules la proporții mai restrînse dă posibilitatea unei investigații statistice mai precise.

In cele ce urmează, considerăm că problema determinării diferitelor caracteristici este rezolvată din punct de vedere practic de îndată ce am determinat distribuțiile lor teoretice. In anumite cazuri, determinăm numai valorile asimptotice ale acestor distribuții, ceea ce are drept consecință că metoda selecției nu se poate aplica pentru ele decît pentru valorile mari ale lui n. In alte cazuri, dăm repartiția exactă a distribuțiilor ce ne interesează, de unde rezultă posibilitatea de a trage concluzii pentru orice valori ale lui n.

Ne vom limita la rezolvarea matematică a problemelor puse de metoda selecției care, de altfel, reprezintă nu numai fundamentul, dar și partea cea mai delicată a acestei teorii.

§ 2. Estimații corecte și absolut corecte

Am văzut că metoda selecției constă în evaluarea caracteristicilor teoretice necunoscute ale colectivității totale, cu ajutorul datelor rezultate dintr-un număr finit de experiențe.

De obicei se consideră că valorile $x_1, x_2, \ldots, x_n$ reprezintă coordonatele unui punct $P(x_1, x_2, \ldots, x_n)$ dintr-un spațiu euclidian n-dimensional.

Fie $F(P)$ o funcție de observațiile date de selecție. Dacă

$$F(P) \underset{n\to\infty}{\to} A,$$

convergența avînt loc *în probabilitate*, $F(P)$ reprezintă o *estimație* a valorii necunoscute A.

Faptul că convergența în medie pătratică implică convergența în probabilitate prezintă mare importanță în acest fel de probleme, deoarece, pentru a conchide că $F(P)$ este o estimație, este suficient să calculăm $M[F(P)]$ și $D^2[F(P)]$. Nu insistăm asupra acestei definiții. Din cauza convergenței în probabilitate, cînd numărul observațiilor este mare, $F(P)$ este foarte apropiat de A, diferența dintre $F(P)$ și A fiind din ce în ce mai mică, atunci cînd $n\to\infty$. Valoarea lui $F(P)$ poate fi, deci, efectiv considerată ca o estimație a lui A. Există, însă, o infinitate de funcții $F(P)$ care tind în probabilitate către A. Vom spune că ele reprezintă estimații corecte ale lui A.

Dacă

$$M[F(P)]=A$$

și

$$D^2[F(P)]\to 0,$$

atunci $F(P)$ tinde în probabilitate către A.

Dacă

$$M[F(P)]=A+B(n)$$

și $B(n)\to 0$, cînd $n\to\infty$, convergența în probabilitate a lui $F(P)$ către A se păstrează cînd $n\to\infty$.

In primul caz, $F(P)$ este o *estimație absolut corectă* a lui A, iar în al doilea caz, o *estimație corectă*. Din cauza termenului $B(n)$ — care se anulează cînd $n\to\infty$, dar pentru n finit, oricît de mare, are o valoare diferită de zero — estimația comportă o eroare sistematică. Această eroare sistematică, dată de $B(n)$, este cu atît mai mică, cu cît n este mai mare. In cazul estimației absolut corecte, eroarea sistematică dispare.

Să considerăm cel mai simplu exemplu și anume o schemă probabilistică a lui J. Bernoulli cu două stări. In fiecare probă, probabili-

Scanned with OKEN Scanner

tatea de a avea o stare este p, iar probabilitatea contrarie q. In n probe consecutive s-a obținut frecvența $\frac{\alpha}{n}$. Se pune problema estimației lui p cu ajutorul lui α.

In acest caz, avem

$$M\left(\frac{\alpha}{n}\right)=p, \quad D^2\left(\frac{\alpha}{n}\right)=\frac{pq}{n}.$$

Rezultă că $\frac{\alpha}{n}$ este o estimație absolut corectă a lui p.

In plus,

$$M\left(\frac{\alpha}{n+1}\right)=p, \quad D^2\left(\frac{\alpha}{n}\right)=\frac{pq}{n}.$$

Prin urmare și $\frac{\alpha}{n+1}$ converge în probabilitate către p. Insă, din cauza termenului $\left(-\frac{p}{n+1}\right)$, valoarea $\frac{\alpha}{n+1}$ este numai o estimație corectă a lui p, comportînd eroarea sistematică negativă $\left(-\frac{p}{n+1}\right)$.

Să estimăm pe p^2. Deoarece am găsit pentru p estimația absolut corectă $\frac{\alpha}{n}$, este natural să încercăm să estimăm pe p^2 prin $\frac{\alpha^2}{n^2}$. Atunci, vom obține

$$M\left(\frac{\alpha^2}{n^2}\right)=p^2+\frac{pq}{n}, \quad D^2\left(\frac{\alpha^2}{n^2}\right)=\frac{4p^3q}{n}+O\left(\frac{1}{n}\right),$$

de unde urmează că $\frac{\alpha^2}{n^2}$ este o estimație corectă pentru p^2.

Pentru a găsi o estimație absolut corectă, observăm că

$$M\left(\frac{\alpha^2}{n^2}\right)=\frac{n-1}{n}p^2+\frac{p}{n}, \quad M\left(\frac{\alpha}{n^2}\right)=\frac{p}{n}.$$

Prin urmare, deducem

$$M\left(\frac{\alpha^2}{n^2}-\frac{\alpha}{n^2}\right)=\frac{n-1}{n}p^2,$$

$$\frac{n}{n-1}M\left(\frac{\alpha^2}{n^2}-\frac{\alpha}{n^2}\right)=p^2,$$

$$M\left(\frac{\alpha^2-\alpha}{n(n-1)}\right)=p^2.$$

Scanned with OKEN Scanner

Cum, însă,

$$D^2\left(\frac{\alpha^2-\alpha}{n(n-1)}\right)=\frac{4p^3q}{n}+O\left(\frac{1}{n^2}\right),$$

reiese că $\frac{\alpha^2-\alpha}{n(n-1)}$ este o estimație absolut corectă a lui p^2.

§ 3. Media de selecție

1. Să considerăm n experiențe independente, reprezentate de selecția

$$x_1,\ x_2,\ \ldots,\ x_n.$$

Atunci media de selecție, dată de expresia

$$\bar{x}=\frac{x_1+x_2+\ldots+x_n}{n},$$

este, evident, o funcție de variabilele $x_1,\ x_2,\ldots,\ x_n$.

Fie $F(x)$ funcția de repartiție a distribuției teoretice care, presupunem, are primele două momente finite. Aceasta înseamnă că fiecare dintre variabilele $x_i\ (i=1,\ 2,\ \ldots,\ n)$ corespunde unei distribuții identice cu distribuția teoretică. Ne propunem să determinăm momentele mediei de selecție $\bar{x}$ în funcție de momentele teoretice

$$M_k=\int_{-\infty}^{+\infty}x^k\,dF(x).$$

Astfel, vom avea

$$\overline{M}_1=M(\bar{x})=M\left(\frac{\sum_{i=1}^{n}x_i}{n}\right)=\frac{1}{n}\sum_{i=1}^{n}M(x_i)=M_1.$$

In același timp, vom obține

$$\overline{M}_2=M(\bar{x}^2)=M\left(\frac{\sum_{i=1}^{n}x_i}{n}\right)^2=\frac{1}{n^2}\sum_{i=1}^{n}M(x_i^2)+\frac{2}{n^2}\sum_{i\neq j}^{n}M(x_i\,x_j)=$$

$$=\frac{M_2}{n}+\frac{n-1}{n}M_1^2.$$

Dispersia mediei de selecție rezultă imediat. In adevăr, putem scrie

$$\overline{m}_2=D^2(\bar{x})=\overline{M}_2-\overline{M}_1^2=\frac{M_2}{n}+\frac{n-1}{n}M_1^2-M_1^2=\frac{M_2-M_1^2}{n}=\frac{\sigma^2}{n},$$

de unde urmează că media de selecție reprezintă o estimație absolut corectă pentru media teoretică. După cum se vede, această proprietate este generală, ea neimplicînd decît existența primelor două momente ale distribuției teoretice.

In mod analog, putem să determinăm și momentele centrate de ordin superior ale mediei de selecție $\bar{x}$. Valoarea exactă a momentului centrat de ordinul r al mediei de selecție $\bar{x}$ este

$$\overline{m}_r = \sum \frac{r!}{j_2!\, j_3! \ldots j_r!} \frac{n_j}{n^r} \left(\frac{m_2}{2!}\right)^{j_2} \left(\frac{m_3}{3!}\right)^{j_3} \ldots \left(\frac{m_r}{r!}\right)^{j_r}, \qquad \text{(XII.3,1)}$$

suma fiind extinsă la toți indicii $j_2, j_3, \ldots, j_r$ întregi, pozitivi sau nuli, supuși condiției $2j_2+3j_3+\ldots+rj_r=r$ și notînd

$$j=j_2+j_3+\ldots+j_r$$

$$n_j = \begin{cases} n(n-1)\ldots(n-j+1) & (j=1,\ 2,\ldots,\ n), \\ 0 & (j=n+1,\ n+2,\ldots). \end{cases}$$

In particular, pentru primele momente ale mediei de selecție $\bar{x}$, deducem din (XII.3,1)

$$\overline{m}_3 = \frac{m_3}{n^2},$$

$$\overline{m}_4 = \frac{3m_2^2}{n^2} + \frac{m_4 - 3m_2^2}{n^3}.$$

. . . . . .

2. Vom cerceta, acum, distribuția asimptotică a mediei de selecție. Aceasta corespunde uneia dintre problemele fundamentale ale calculului probabilităților, cunoscută sub numele de *teorema centrală*, enunțată încă de Laplace în condiții foarte puțin restrictive. Problema este azi complet rezolvată, în sensul că s-au determinat condițiile necesare și suficiente pentru ca distribuția mediei de selecție să urmeze la limită o distribuție normală. Aici, ne vom mărgini să arătăm numai condițiile suficiente, foarte generale, date de Liapunov, pentru a obține la limită o distribuție normală. Aceste condiții sînt îndeplinite în majoritatea problemelor întîlnite în statistica matematică.

Scanned with OKEN Scanner

Să considerăm şirul de variabile independente

$$y_1, y_2, \ldots, y_n,$$

unde fiecare dintre aceste variabile corespunde unei distribuţii bine determinate, dar care poate să ne fie necunoscută, şi să presupunem numai că există primele trei momente ale distribuţiei fiecăreia dintre aceste variabile, adică

$$\begin{aligned} m_k &= M(y_k), \\ \sigma_k^2 &= M(y_k - m_k)^2, \\ \rho_k^3 &= M(|y_k - m_k|)^3. \end{aligned} \qquad \text{(XII.3,2)}$$

Atunci, dacă

$$\lim_{n\to\infty} \frac{\left(\sum_{k=1}^{n} \rho_k^3\right)^{\frac{1}{3}}}{\left(\sum_{k=1}^{n} \sigma_k^2\right)^{\frac{1}{2}}} = 0, \qquad \text{(XII.3,3)}$$

teorema lui Liapunov afirmă că distribuţia sumei de variabile independente

$$y_1 + y_2 + \ldots + y_n$$

tinde la limită către distribuţia normală.

Notînd $y_k = \frac{x_k}{n}$, media de selecţie $\bar{x}$ devine

$$\bar{x} = y_1 + y_2 + \ldots + y_n.$$

In acest caz, primele trei momente (XII.3,2) sînt

$$\begin{aligned} m_k &= M\left(\frac{x_k}{n}\right) = \frac{M_1}{n}, \\ \sigma_k^2 &= M\left(\frac{x_k - M_1}{n}\right)^2 = \frac{\sigma^2}{n^2}, \\ \rho_k^3 &= M\left(\frac{|x_k - M_1|}{n}\right)^3 = \frac{\rho^3}{n^3} \end{aligned}$$

și, deci, condiția (XII.3,3), la rîndul ei, devine

$$\lim_{n\to\infty}\frac{\left(\sum_{k=1}^{n}\rho_k^3\right)^{\frac{1}{3}}}{\left(\sum_{k=1}^{n}\sigma_k^2\right)^{\frac{1}{2}}}=\lim_{n\to\infty}\frac{\rho}{n^{\frac{2}{3}}}\cdot\frac{n^{\frac{1}{2}}}{\sigma}=\lim_{n\to\infty}\frac{\rho}{\sigma n^{\frac{1}{6}}}=0,$$

dacă $\sigma\neq 0$ și ρ este finit.

Prin urmare, distribuția mediei de selecție tinde la limită către distribuția normală, cu condiția ca să admită momentele m_k, σ_k, ρ_k.

3. Am arătat mai sus că distribuția mediei de selecție, în anumite condiții, se poate determina pentru $n\to\infty$. In unele cazuri particulare, distribuția mediei de selecție se poate determina, chiar pentru valorile finite ale lui n numărul probelor de selecție. Teoremele următoare confirmă cele afirmate mai sus.

Teoremă. *Dacă distribuția teoretică este normală* $N(m, \sigma)$, *distribuția mediei de selecție este de asemenea normală* $N\left(m, \frac{\sigma}{\sqrt{n}}\right)$.

Pentru demonstrarea teoremei, folosim metoda funcției caracteristice. Avem

$$\varphi(t)=M(e^{it\bar{x}})=M\left(e^{i\frac{t}{n}\sum_{k=1}^{n}x_k}\right)=\prod_{k=1}^{n}M\left(e^{i\frac{t}{n}x_k}\right).$$

Efectuînd substituția $u=\frac{t}{n}$, funcția caracteristică devine

$$\varphi(u)=\prod_{k=1}^{n}M(e^{iux_k}). \qquad \text{(XII.3,4)}$$

Insă, din studiul distribuției normale de o singură variabilă (cap. VI, § 2) se știe că

$$M(e^{iux_k})=e^{mui-\frac{\sigma^2u^2}{2}},$$

de unde urmează că funcția caracteristică (XII.3,4) este

$$\varphi(t)=e^{nmui-\frac{n\sigma^2u^2}{2}}=e^{mti-\frac{\sigma^2u^2}{2n}},$$

ținînd seama de substituția $u=\frac{t}{n}$.

După cum se vede, aceasta este funcția caracteristică a unei distribuții normale $N\left(m, \frac{\sigma}{\sqrt{n}}\right)$. Atunci și funcția de repartiție corespunzătoare este normală $N\left(m, \frac{\sigma}{\sqrt{n}}\right)$ și teorema este demonstrată.

Teoremă. *Dacă distribuția teoretică este o distribuție Poisson cu parametrul λ, atunci distribuția mediei de selecție are ca funcție caracteristică expresia*

$$\varphi(t) = e^{n\lambda\left(e^{i\frac{t}{n}}-1\right)}.$$

Folosind din nou metoda funcției caracteristice, avem

$$\varphi(t) = M(e^{it\bar{x}}) = \prod_{k=1}^{n} M\left(e^{i\frac{t}{n}x_k}\right)$$

sau, în urma substituției $u = \frac{t}{n}$,

$$\varphi(u) = \prod_{k=1}^{n} M(e^{iux_k}). \qquad \text{(XII.3,5)}$$

Avînd în vedere funcția caracteristică a distribuției Poisson cu parametrul λ, expresia (XII.3,5) devine

$$\varphi(u) = \prod_{k=1}^{n} e^{\lambda(e^{iu}-1)} = e^{n\lambda(e^{iu}-1)},$$

iar în baza aceleiași substituții $u = \frac{t}{n}$ devine

$$\varphi(t) = e^{n\lambda\left(e^{i\frac{t}{n}}-1\right)},$$

ceea ce demonstrează teorema.

Observăm că distribuția cu această funcție caracteristică este o distribuție apropiată de distribuția Poisson cu parametrul λ.

Teormă. *Dacă distribuția teoretică este o distribuție Cauchy, distribuția mediei de selecție urmează de asemenea o distribuție Cauchy.*

Procedînd ca și în teoremele precedente, prin metoda funcției caracteristice, avem

$$\varphi(t) = M(e^{it\bar{x}}) = \prod_{k=1}^{n} M\left(e^{i\frac{t}{n}x_k}\right),$$

adică

$$\varphi(u)=\prod_{k=1}^{n} M(e^{iux})=e^{n|u|},$$

făcînd transformarea $u=\frac{t}{n}$ și reamintind că funcția caracteristică a distribuției Cauchy este $e^{|u|}$.
Revenind la expresia funcției caracteristice în t, rezultă

$$\varphi(t)=e^{n\left|\frac{t}{n}\right|}=e^{|t|}.$$

Prin urmare, teorema este demonstrată.
Acesta constituie un caz mai special, cînd teorema limită centrală nu este satisfăcută. In acest caz, condițiile lui Liapunov nu sînt satisfăcute deoarece, după cum am arătat, distribuția Cauchy nu are momente de nici un ordin.

Teoremă. *Dacă distribuția teoretică este o distribuție rectangulară din intervalul* [0, 1], *atunci media de selecție admite o distribuție a cărei funcție de frecvență este*

$$f(x)=\frac{n}{(n-1)!}\sum_{h=0}^{k\leq nx}(-1)^k C_n^k (nx-k)^{n-1}.$$

Plecăm, pentru demonstrație, de la formula de inversiune a lui Fourier, care exprimă funcția de frecvență prin funcția caracteristică a distribuției și anume

$$f(x)=\frac{1}{2\pi}\int_{-\infty}^{+\infty} e^{-itx}\varphi(t)\,dt. \qquad \text{(XII.3,6)}$$

Știind că funcția caracteristică a distribuției teoretice în cazul de față este dată de expresia

$$\int_0^1 e^{itx}dx=\frac{e^{it}-1}{it},$$

atunci funcția caracteristică, corespunzătoare distribuției mediei de selecție $\bar{x}$, este

$$\varphi(t)=M(e^{it\bar{x}})=\prod_{k=1}^{n} M\left(e^{i\frac{t}{n}x_k}\right)=\left(\frac{e^{i\frac{t}{n}}-1}{i\frac{t}{n}}\right)^n.$$

Inlocuind această expresie a lui $\varphi(t)$ în (XII.3,6), obținem

$$f(x)=\frac{1}{2\pi}\int_{-\infty}^{+\infty}e^{-itx}\left(\frac{e^{i\frac{t}{n}}-1}{i\frac{t}{n}}\right)^n dt=$$

$$=\frac{n^n}{2\pi i^n}\int_{-\infty}^{+\infty}e^{itx}\sum_{k=0}^{n}(-1)^{n-k}C_n^k\frac{e^{\frac{itk}{n}}}{t^n}dt=$$

$$=\frac{n^n}{2\pi i^n}\sum_{k=0}^{n}(-1)^{n-k}C_n^k\int_{-\infty}^{+\infty}\frac{e^{i\left(\frac{k}{n}-x\right)t}}{t^n}dt. \qquad \text{(XII.3,7)}$$

Se știe că o integrală de forma

$$\int_{-\infty}^{+\infty}\frac{e^{iax}}{x^n}dx=\begin{cases}-2\pi i^n\dfrac{a^{n-1}}{(n-1)!} & (a\leq 0)\\ 0 & (a>0).\end{cases}$$

Prin urmare, expresia (XII.3,7) devine

$$f(x)=\frac{n^n}{(n-1)!}\sum_{k=0}^{k\leq nx}(-1)^{n-k+1}C_n^k\left(\frac{k}{n}-x\right)^{n-1}=$$

$$=\frac{(-1)^{n-1}n^n}{(n-1)!}\sum_{k=0}^{k\leq nx}(-1)^{n-k+1}C_n^k\left(x-\frac{k}{n}\right)^{n-1}=$$

$$=\frac{(-1)^n n}{(n-1)!}\sum_{k=0}^{k\leq nx}(-1)^{n-k+1}C_n^k(nx-k)^{n-1}=$$

$$=\frac{n}{(n-1)!}\sum_{k=0}^{k\leq nx}(-1)^{-k}C_n^k(nx-k)^{n-1}=$$

$$=\frac{n}{(n-1)!}\sum_{k=0}^{k\leq nx}(-1)^{k}C_n^k(nx-k)^{n-1},$$

deoarece $(-1)^{-k}=\dfrac{1}{(-1)^k}=(-1)^k$, oricare ar fi numărul k întreg pozitiv sau nul.

Teorema este demonstrată.

§ 4. Momente de selecție

1. Momentul de selecție de ordinul k definit de expresia

$$\overline{M}_k = \frac{x_1^k + x_2^k + \ldots + x_n^k}{n}$$

este, după cum se vede, o variabilă aleatorie. Atunci, avem

$$M(\overline{M}_k) = M\left(\frac{\sum_{i=1}^{n} x_i^k}{n}\right) = \frac{\sum_{i=1}^{n} M(x_i^k)}{n} = M_k,$$

$$M(\overline{M}_k - M_k)^2 = \frac{1}{n^2} M\left[\sum_{i=1}^{n} (x_i^k - M_k)\right]^2 =$$

$$= \frac{1}{n^2} \sum_{i=1}^{n} M(x_i^k - M_k)^2 = \frac{M_{2k} - M_k^2}{n}.$$

Cu alte cuvinte, momentul de selecție de ordinul k, $\overline{M}_k$ converge în probabilitate către momentul teoretic de ordinul k, M_k cu condiția ca momentele teoretice M_k și M_{2k} să existe. Deoarece $\overline{M}_k$ se prezintă ca o sumă de variabile independente $y_i = \frac{x_i}{n}$ $(i = 1, 2, \ldots, n)$, în baza condițiilor necesare și suficiente stabilite de Liapunov, distribuția lui $\overline{M}_k$ tinde la limită către distribuția normală $N\left(M_k, \sqrt{\frac{M_{2k} - M_k^2}{n}}\right)$ dacă, în afară de existența momentelor M_k și M_{2k} mai există și momentul $M(|\overline{M}_k - M_k|)^3$.

2. Se pune, acum, problema determinării momentelor de un ordin oarecare ale distribuției lui $\overline{M}_k$. Această problemă a fost rezolvată de T. Lișcu, care a stabilit expresia exactă a acestor momente. Astfel, pentru momentul de ordinul r, avem următoarea expresie

$$M(\overline{M}_k^r) = \sum \frac{r!}{j_1!\, j_2! \ldots j_r!} \frac{n_J}{n^r} \left(\frac{M_k}{1!}\right)^{j_1} \left(\frac{M_{2k}}{2!}\right)^{j_2} \cdots \left(\frac{M_{rk}}{r!}\right)^{j_r},$$

suma fiind extinsă la toți indicii $j_1, j_2, \ldots, j_r$, întregi pozitivi sau nuli, supuși condiției

$$j_1+2j_2+\ldots+rj_r=r$$

și notînd

$$j=j_1+j_2+\ldots+j_r$$

$$n_j=\begin{cases} n(n-1)\ldots(n-j+1) & (j=1, 2, \ldots, n) \\ 0 & (j=n+1,\ n+2\ldots). \end{cases}$$

De asemenea, T. Lișcu a stabilit și covarianța între două momente de selecție $\overline{M}_k^r$ și $\overline{M}_{k'}^{r'}$, a cărei expresie este

$$M(\overline{M}_r^k,\ \overline{M}_{kl}^{r'}) =$$

$$=\sum \frac{r!\,r'!}{j_{10}!\,j_{01}!\ldots j_{rr}!}\cdot\frac{n_j}{n^{r+r'}}\left(\frac{M_k}{1!\,0!}\right)^{j_{10}}\left(\frac{M_{k'}}{0!\,1!}\right)^{j_{01}}\cdots\left(\frac{M_{rk+r'k'}}{r!\,r'!}\right)^{j_{rr'}},$$

suma fiind extinsă la toți indicii $j_{10}, j_{01}, \ldots, j_{rr'}$, întregi pozitivi sau nuli, supuși condițiilor

$$\sum_{i,\,i'} ij_{ii'}=r,\quad \sum i'j_{ii'}=r'\quad (i=0, 1, \ldots, r),\quad (i'=0, 1, \ldots, r')$$

și notînd

$$j=\sum_{i,\,i'} j_{i,\,i'},$$

$$n_j=\begin{cases} n(n-1)\ldots(n-j+1) & (j=1, 2, \ldots, n) \\ 0 & (j=n+1,\ n+2, \ldots)' \end{cases}$$

$0!=1.$

Aceste formule rezultă ca un caz particular dintr-o formulă mai generală. Considerînd media

$$\overline{u}=\frac{1}{n}\sum_{i=1}^{n} u(x_i),$$

unde $x_i\ (i=1, 2, \ldots, n)$ sînt variabile independente, T. Lișcu a reușit să calculeze momentul de ordinul r, $M(\overline{u}^r)$, în funcție de momentele variabilei $u(x)$. Demonstrația lui pornește de la o generalizare

a formulei de derivare dată de Faa de Bruno, pe care o redăm mai jos.

Derivata parțială de ordinul r în raport cu x și de ordinul r' în raport cu x' a unei funcții de funcție de x și x'

$$\varphi(x, x') = F(f(x, x'))$$

are ca expresie

$$\varphi_{rr'} = \sum \frac{r!\, r'!}{j_{10}!\, j_{01}! \ldots j_{rr'}!} F^{(j)} \left(\frac{f_{10}}{1!\, 0!}\right)^{f_{10}} \left(\frac{f_{01}}{0!\, 1!}\right)^{f_{01}} \cdots \left(\frac{f_{rr'}}{r!\, r'!}\right)^{f_{rr'}},$$

suma fiind extinsă la toți indicii $j_{10}, j_{01}, \ldots, j_{rr'}$, întregi pozitivi sau nuli, supuși condițiilor

$$\sum_{i,i'} i j_{ii'} = r, \quad \sum_{i,i'} i' j_{ii'} = r' \quad (i = 0, 1, \ldots, r) \ (i' = 0, 1, \ldots, r')$$

și notînd

$$j = \sum_{i,i'} j_{ii'}, \quad 0! = 1.$$

In cele ce urmează, vom arăta pe scurt cum se ajunge la această formulă.

Fie

$$Y = a_{10}X + a_{01}X' + a_{20}X^2 + a_{11}XX' + a_{02}X'^2 + \ldots$$
$$\ldots + a_{0s}X'^s = \sum_{i,i'} a_{ii'} X^i X'^{i'}$$
$$(i,\ i' = 0, 1, \ldots, s) \ \text{(exceptînd } i = i' = 0)$$
$$Z = b_1 Y + b_2 Y^2 + \ldots + b_q Y^q = C_{10}X + C_{01}X' + \quad \text{(XII.4,1)}$$
$$+ C_{20}X^2 + C_{11}XX' + C_{02}X'^2 + \ldots$$

Coeficientul lui $X^r X'^{r'}$ în dezvoltarea lui Y^j este

$$\sum \frac{j!}{j_{10}!\, j_{01}! \ldots j_{os}!} a_{10}^{j_{10}} a_{01}^{j_{01}} \ldots a_{0s}^{j_{0s}},$$

astfel ca

$$\sum_{i,i'} j_{ii'} = j, \quad \sum_{i,i'} i j_{i,i'} = r, \quad \sum_{i,i'} i' j_{ii'} = r'$$
$$(i, i' = 0, 1, \ldots, s) \qquad \text{(exceptînd } i = i' = 0).$$

Scanned with OKEN Scanner

Penultima condiție exclude valorile lui i mai mari decît r, ultima condiție exclude valorile lui i' mai mari decît r', deci rămîne

$$(i=0,\ 1,\ \ldots,\ r), \qquad (i'=0,\ 1,\ \ldots,\ r').$$

Apoi $X^r X^{r'}$ nu poate figura în dezvoltările puterilor lui Y mai mari decît $r+r'$, deci rămîne

$$c_{rr'} = \sum_{j=1}^{r+r'} b_j \sum \frac{j!}{j_{10}!\ j_{01}!\ldots j_{rr'}!}\, a_{10}^{j_{10}} a_{01}^{j_{01}} \ldots a_{rr'}^{j_{rr'}}. \qquad \text{(XII.4,2)}$$

Fie

$$y = f(x, x')$$

o funcție parțial derivabilă de s ori în raport cu x și în raport cu x',

$$z = F(y)$$

o funcție derivabilă de q ori,

$$z = F(f(x, x')) = \varphi(x, x'),$$

$$y_0 = f(x_0, x_0'), \quad z_0 = F(y_0).$$

Să introducem notațiile

$$x - x_0 = X, \quad x' - x_0' = X', \quad y - y_0 = Y, \quad z - z_0 = Z,$$

$$f_{ii'}(x_0, x_0') = i!\, i'!\, a_{ii'} \qquad (i+i'=0,\ 1,\ \ldots,\ s-1),$$

$$f_{ii'}(x_0+\theta x, x_0'+\theta x') = i!\, i'!\, a_{ii'} \qquad (i+i'=s) \qquad 0<\theta<1.$$

$$F^{(j)}(y_0) = j!\, b_j, \quad (j=0, 1, \ldots, q-1),$$

$$F^{(q)}(y_0+\theta' Y) = q!\, b_q, \qquad 0<\theta'<1.$$

Cu această notație, formula lui Taylor, aplicată funcțiilor $f(x, x')$, $F(y)$ ne dă expresiile (XII.4,1).
Pentru $r+r' < {}^{s}_{q}$, formula (XII.4,2) se scrie

$$c_{rr'} = \sum_{j=1}^{r+r'} \frac{1}{j!} F^{(j)}(y_0) \sum \frac{j!}{j_{10}!\ j_{01}!\ldots j_{rr'}!} \left(\frac{f_{10}(x_0, x_0')}{1!\ 0!}\right)^{j_{10}} \ldots \left(\frac{f_{rr'}(x_0, x_0')}{r!\ r'!}\right). \qquad \text{(XII.4,3)}$$

Deci $c_{rr'}$ nu depinde de x și x'.

Scanned with OKEN Scanner

Atunci, din (XII.4,1) rezultă prin derivare

$$c_{rr'}=\frac{1}{r!\,r'!}\,\varphi_{rr'}(x_0, x_0'). \qquad \text{(XII.4,4)}$$

Suprimînd în (XII.4,3) și în (XII.4,4) indicele 0 de la x_0 și x_0' obținem formula din enunț.

Formula se extinde imediat pentru mai multe variabile și, restrînsă pentru o singură variabilă, redă formula lui Faa de Bruno.

Derivata de ordinul r a unei funcții de funcție $\varphi(x)=F(f(x))$ are ca expresie

$$\varphi(r)=\sum \frac{r!}{j_1!\,j_2!\ldots j_r!}\,F^{(j)}\left(\frac{f'}{1!}\right)^{j_1}\left(\frac{f''}{2!}\right)^{j_2}\cdots\left(\frac{f^{(r)}}{r!}\right)^{j_r},$$

suma fiind extinsă la toți indicii $j_1, j_2, \ldots, j_r$ întregi pozitivi sau nuli, supuși condiției

$$j_1+2j_2+\ldots+rj_r=r$$

și notînd

$$j=j_1+j_2+\ldots+j_r.$$

Conform relației

$$\varphi(t, t')=\left(\psi\left(\frac{t}{n}, \frac{t'}{n}\right)\right)^n,$$

unde $\varphi(t, t')$ este funcția generatoare a momentelor $M(\overline{u}^r, \overline{u}'^{r'})$, iar $\varphi(t, t')$ funcția generatoare a momentelor $M(u^h, u'^h)$, $\varphi(t, t')$ este o funcție de funcție de t și t'

$$\varphi(t, t')=F\left(\psi\left(\frac{t}{n}, \frac{t'}{n}\right)\right),$$

știind că

$$F(y)=y^n.$$

Rezultă

$$\left(\psi\left(\frac{t}{n}, \frac{t'}{n}\right)\right)_{hh'}=\frac{1}{n^{h+h'}}\,\psi_{hh'}\left(\frac{t}{n}, \frac{t'}{n}\right) \qquad (h, h'=0, 1, \ldots),$$

$$F^{(j)}(y)=\begin{cases} n(n-1)\ldots(n-j+1)\,y^{n-j} & (j=1, 2, \ldots, n) \\ 0 & (j=n+1, n+2, \ldots). \end{cases}$$

Pentru $t=t'=0$, obținem

$$\psi(0,\ 0)=1,\quad \left[\psi\left(\frac{t}{n},\ \frac{t'}{n}\right)\right]_{t=t'=0}=\frac{1}{n^{h+h'}}M(u^h u'^{h'}),$$

$$F(1)\ =1,\quad F^{(j)}(1)=n_j,$$

$$\varphi(0,\ 0)=1,\quad \varphi_{rr'}(0,\ 0)=M(\overline{u}^r u'^{r'}).$$

Scriind, o dată cu aceste expresii, formula derivatei parțiale de ordinul r în raport cu t și de ordinul r' în raport cu t' a unei funcții de funcție de t și t' pentru $t=t'=0$, obținem formulele de mai jos

$$M(\overline{u}^r)=\sum\frac{r!}{j_1!\,j_2!\ldots j_r!}\cdot\frac{n_j}{n^r}\left(\frac{M(u)}{1!}\right)^{j_1}\left(\frac{M(u^2)}{2!}\right)^{j_2}\ldots\left(\frac{M(u^r)}{r!}\right)^{j_r},$$

suma fiind extinsă la toți indicii $j_1, j_2, \ldots, j_r$ întregi pozitivi sau nuli, supuși condiției

$$j_1+2j_2+\ldots+rj_r=r$$

și notînd

$$j=j_1+j_2+\ldots+j_r$$

$$n_j=\begin{cases} n(n-1)\ldots(n-j+1) & (j=1,\ 2,\ \ldots,\ n)\\ 0 & (j=n+1,\ n+2\ldots),\end{cases}$$

$$M(\overline{u}^r u'^{r'})=$$

$$=\sum\frac{r!\,r'!}{j_{10}!\,j_{01}!\ldots j_{rr'}!}\cdot\frac{u_j}{n^{r+r'}}\left(\frac{M(u)}{1!\,0!}\right)^{j_{10}}\left(\frac{M(u')}{0!\,1!}\right)^{j_{01}}\ldots\left(\frac{M(u^r u'^{r'})}{r!\,r'!}\right)^{j_{rr'}},$$

suma fiind extinsă la toți indicii $j_{10}, j_{01}, \ldots, j_{rr'}$, întregi pozitivi sau nuli, supuși condițiilor

$$\sum_{i,i'} ij_{ii'}=r,\quad \sum_{i,i'} i'j_{ii'}=r' \qquad (i=0,\ 1,\ \ldots,\ r),\ (i'=0,\ 1,\ \ldots,\ r')$$

și notînd

$$j=\sum_{i,i'} j_{ii'},$$

$$n_j=\begin{cases} n(n-1)\ldots(n-j+1) & (j=1,\ 2,\ \ldots,\ n)\\ 0 & (j=n+1,\ n+2\ldots)\end{cases},$$

$$0!=1.$$

Extinderea formulelor pentru mai multe variabile este imediată. Prima formulă rezultă din a doua pentru $r'=0$.

Particularizînd funcțiile u și u'

$$u=x^k, \quad u'=x^{k'},$$

în care caz

$$\overline{u}=\overline{M}_k, \quad \overline{u}'=\overline{M}_{k'}$$

și scriind expresia momentelor $M(\overline{u}^r)$ și $M(\overline{u}^r u'^{r'})$ obținem expresiile momentelor $M(\overline{M}_k^r)$ și $M(\overline{M}_k^r \overline{M}_{k'}^{r'})$, pe care le-am scris mai înainte.

Tot din aceste formule generale rezultă și expresia momentului centrat de ordinul r al mediei de selecție pe care am dat-o în paragraful precedent. In adevăr, pentru $p=1$, în care caz

$$M_1=\overline{x},$$

se obține expresia momentului $M(\overline{x}^r)$ al mediei de selecție $\overline{x}$.

In sfîrșit, fie

$$u=x-M_1,$$

atunci

$$\overline{u}=\overline{x}-M_1.$$

Scriind expresia momentului $M(\overline{u}^r)=M(\overline{x}-M_1)^r$, care este momentul centrat de ordinul r al mediei de selecție x în funcție de momentele

$$M(u^h)=M(x-M_1)^h=m_h \qquad (h=2,\ 3,\ \ldots,\ r)$$

se obține, după cum am spus, expresia momentului centrat de ordinul r al mediei de selecție $\overline{x}$, $M(\overline{x}-M_1)^r$.

§ 5. Momente centrate de selecție

Momentul centrat de selecție de ordinul k, definit de expresia

$$\overline{m}_k=\frac{(x_1-\overline{x})^k+(x_2-\overline{x})^k+\ldots+(x_n-\overline{x})^k}{n},$$

este o funcție de $x_1, x_2, \ldots, x_n$ și, deci, este o variabilă aleatorie. Determinarea momentelor distribuției lui $\overline{m}_k$ dă naștere la calcule

mai complicate decît cele întîlnite la momentele distribuției lui $\overline{M}_k$, din cauză că abaterile $x_i - \overline{x}$ $(i = 1, 2, \ldots, n)$ se raportează la $\overline{x}$, care la rîndul său este o variabilă aleatorie. Pentru simplificarea calculelor, vom presupune în cele ce urmează că $m = 0$, ceea ce nu restrînge generalitatea, mai cu seamă că m_k rămîne, după cum s-a arătat mai înainte (cap. II, § 3), invariant la o schimbare a originii.

Teoremă. *Valoarea medie a distribuției variabilei $\overline{m}_k$, momentul centrat de selecție de ordin k, se exprimă cu ajutorul formulei*

$$M(\overline{m}_k) = m_k + o\left(\frac{1}{n}\right),$$

unde $o\left(\frac{1}{n}\right)$ este o funcție asimptotică de ordinul lui $\frac{1}{n}$.

Să observăm că

$$(x_i - \overline{x})^k = x_i^k - C_k^1 \overline{x} x_i^{k-1} + C_k^2 \overline{x}^2 x_i^{k-2} - \ldots$$

Rezultă că

$$\overline{m}_k = \overline{M}_k - C_k^1 \overline{x} \overline{M}_{k-1} + C_k^2 \overline{x}^2 \overline{M}_{k-2} - \ldots$$

Să mai observăm că $M_k = m_k$ pentru orice k, deoarece am presupus $m = 0$. Pe de altă parte

$$M(\overline{x}^s \overline{M}_{k-s}) \leq \sqrt{M(\overline{x}^{2s})\, M(\overline{M}_{k-s}^2)},$$

conform inegalității lui Schwarz.

Am arătat, însă, că

$$M(\overline{M}_{k-s}^2) = M_{k-s}^2,$$

$$M(\overline{x}^{2s}) = o\left(\frac{1}{n^s}\right).$$

Prin urmare, expresia $M(\overline{x}^s \overline{M}_{k-s})$ este asimptotică cel mult de ordinul lui $\frac{1}{n^s}$ pentru orice s.

Din expresia lui $\overline{m}_k$ rezultă, atunci, imediat

$$M(\overline{m}_k) = m_k + o\left(\frac{1}{n}\right),$$

ceea ce demonstrează teorema.

Scanned with OKEN Scanner

Teoremă. *Dispersia distribuției variabilei m_k, momentul centrat de selecție de ordinul k, se exprimă cu formula*

$$D^2(\overline{m}_k) = \frac{m_{2k} - m_k^2 - 2km_{k-1}m_{k+1} + k^2 m_2 m_{k-1}^2}{n} + o\left(\frac{1}{n^2}\right),$$

iar covarianța momentelor $\overline{M}_h$ și $\overline{M}_k$ este dată de formula

$$M(\overline{M}_h - m_h)(\overline{M}_k - m_k) =$$

$$= \frac{m_{h+k} - m_h m_k - hm_{h-1}m_{k+1} - km_{h+1}m_{k-1} + hkm_{h-1}m_{k-1}}{n} + o\left(\frac{1}{n^2}\right),$$

unde $o\left(\frac{1}{n^2}\right)$ este o funcție asimptotică de ordinul lui $\frac{1}{n^2}$.

Atunci, vom avea

$$D^2(\overline{m}_k) = M(\overline{m}_k - m_k)^2 = M(\overline{M}_k - m_k - k\bar{x}\overline{M}_{k-1} + c_k^2 \bar{x}^2 \overline{M}_{k-2} + \ldots)^2 =$$

$$= M(\overline{M}_k - m_k)^2 + k^2 M(\bar{x}^2 \overline{M}_{k-1}^2) - 2kM(\bar{x}\overline{M}_{k-1}(\overline{M}_k - m_k)) + T,$$

unde s-a notat cu T restul termenilor care provin din ridicarea la pătrat a expresiei $(\overline{m}_k - m_k)$.
Insă,

$$M(\overline{M}_k - m_k)^2 = \frac{m_{2k} - m_k}{n},$$

$$M(\bar{x}^2 \overline{M}_{k-1}) = \frac{1}{n^4} M\left(\sum_{i=1}^n x_i \sum_{i=1}^n x_i^{k-1}\right)^2 =$$

$$= \frac{1}{n^4} M\left(\sum_{i=1}^n x_i^2 + 2\sum_{i<j} x_i x_j\right)\left(\sum_{h=1}^n x_h^{2k-2} + 2\sum_{h\neq l}^n x_h^{k-1} x_l^{k-1}\right) =$$

$$= \frac{1}{n^4} M\left(\sum_{i\neq h}^n x_i^2 x_h^{2k-2} + 2\sum_{i\neq h\neq l}^n x_i^2 x_h^{k-1} x_l^{k-1} + \right.$$

$$\left. + 2\sum_{i\neq j\neq h} x_i x_j x_h^{2k-2} + 4\sum_{i\neq j\neq h\neq k} x_i x_j x_h^{k-1} x_k^{k-1}\right).$$

Intrucît $M(x_i)=0$ şi variabilele x_i sînt independente, termenii

$$2\sum_{\substack{i,\,j,\,h=1\\ i\neq j\neq h}}^{n} x_i x_j x_h^{2k-2}, \quad 4\sum_{\substack{i,\,j,\,h,\,k=1\\ i\neq j\neq h\neq k}}^{n} x_i x_j x_h^{k-1} x_k^{k-1}.$$

dau valori medii nule.
Pe de altă parte,

$$M(x_i^2 x_h^{2k-2}) = m_2 m_{2k-2}, \quad M(x_i^2 x_h^{k-1} x_l^{k-1}) = m_2 m_{k-1}^2,$$

pentru orice i, h, l diferiţi între ei.
Rămîne să numărăm cîţi termeni sînt de forma $x_i^2 x_h^{2k-2}$ şi cîţi de forma $x_i^2 x_h^{k-1} x_l^{k-1}$, cînd indicii i, h, l iau în condiţiile de mai sus toate valorile de la 1 la n. Numărul termenilor de forma

$$x_i^2 x_h^{2k-2}, \quad x_i^2 x_h^{k-1} x_l^{k-1}$$

este, respectiv, egal cu

$$n(n-1), \quad \frac{n(n-1)(n-2)}{2}.$$

Prin urmare, valoarea asimptotică a momentului $M(\bar{x}^2 \overline{M}_{k-1}^2)$ va fi dată numai de grupul de termeni

$$2\sum_{i\neq h\neq l} x_i^2 x_h^{k-1} x_l^{k-1},$$

pentru care

$$\frac{2}{n^4} M\left(\sum_{i\neq h\neq l} x_i^2 x_h^{k-1} x_l^{k-1}\right) = \frac{1}{n} m_2 m_{k-1}^2 + o\left(\frac{1}{n^2}\right),$$

$$M(\bar{x}\overline{M}_{k-1}(\overline{M}_k - m_k)) = M(\bar{x}\overline{M}_{k-1}\overline{M}_k) - m_k M(\bar{x}\overline{M}_{k-1}) =$$

$$= \frac{1}{n^3} M\left(\sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i^{k-1} \sum_{i=1}^{n} x_i^k\right) - \frac{m_k}{n^2} M\left(\sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i^{k-1}\right).$$

Efectuînd calculele, vom observa că pentru evaluarea asimptotică a momentului unui produs de forma

$$M\left(\sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i^{k-1} \sum_{i=1}^{n} x_i^k\right)$$

va trebui să păstrăm termenii $x_i^{\alpha_i} x_j^{\alpha_j}, \ldots$, în care intră cei mai mulți indici $i, j, \ldots$, cu excepția valorilor $\alpha_i, \alpha_j, \ldots$, egale cu unu, care conduc la un moment nul.

Rezultă că

$$\frac{1}{n^3} M\left(\sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i^{k-1} \sum_{i=1}^{n} x_i^k\right) = \frac{1}{n^3} M\left(\sum_{i \neq j} x_i^{k-1} x_j^{k-1} + \sum_{i \neq j} x_i^k x_j^k\right)$$

și

$$M(x_i^{k-1} x_j^{k+1}) = m_{k-1} m_{k+1}; \quad M(x_i^k x_j^k) = m_k^2.$$

Avem atunci $n(n-1)$ termeni $x_i^{k-1} x_j^{k+1}$ și $x_i^k x_j^k$. Deci,

$$\frac{1}{n^3} M\left(\sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i^{k-1} \sum_{i=1}^{n} x_i^k\right) = \frac{m_{k-1} m_{k+1} + m_k^2}{n} + o\left(\frac{1}{n^2}\right).$$

De asemenea,

$$\frac{m_k}{n^2} M\left(\sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i^{k-1}\right) = \frac{m_k}{n^2} M\left(\sum_{i=1}^{n} x_i^k\right) = \frac{m_k^2}{n}.$$

Prin urmare, obținem

$$M(\bar{x} \overline{M}_{k-1} (\overline{M}_k - m_k)) = \frac{m_{k-1} m_{k+1}}{n} + o\left(\frac{1}{n}\right).$$

A mai rămas să evaluăm valoarea asimptotică a momentelor date de T.

Procedînd în același mod, se vede că

$$M(\bar{x}^2 \overline{M}_{k-2} (\overline{M}_k - m_k)) = o\left(\frac{1}{n^2}\right).$$

Toți ceilalți termeni, care sînt cuprinși în T, vor avea valori medii de forma

$$\frac{1}{n^r} M(x_i^\alpha x_j^\beta x_h^\gamma x_l^\delta),$$

dacă $r > 5$ și

$$\frac{1}{n^r} M(x_i^\alpha x_j^\beta x_h^\gamma)$$

dacă $r = 5$.

Se vede că asimptotic acești termeni sînt cel puțin de ordinul lui $\frac{1}{n^2}$.

Scanned with OKEN Scanner

In sfîrșit, ținînd seama de valorile medii pe care le-am calculat, rezultă că

$$D^2(\overline{m}_k) = M(\overline{m}_k - m_k)^2 = \frac{m_{2k} - m_k^2 - 2km_{k-1}m_{k+1} + k^2 m_2 m_{k-1}^2}{n} + o\left(\frac{1}{n^2}\right).$$

In același mod, fără a mai insista asupra calculelor, se arată că pentru covarianța momentelor $\overline{M}_h$ și $\overline{M}_k$ ajungem la formula următoare

$$M(\overline{M}_h - m_h)(\overline{M}_k - m_k) =$$

$$= \frac{m_{h+k} - m_h m_k - hm_{h-1}m_{k+1} - km_{h+1}m_{k-1} + hkm_{h-1}m_{k-1}}{n} + o\left(\frac{1}{n^2}\right).$$

Teorema este, astfel, demonstrată.

2. **Teoremă.** *Momentul centrat de selecție de ordinul k, $\overline{m}_h$ este o estimație corectă a lui m_k.*

Demonstrația teoremei rezultă imediat.

In adevăr, din teoremele precedente, care ne arată valoarea medie și dispersia distribuției $\overline{m}_k$, se vede că $\overline{m}_k$ este o estimație corectă a lui m_k.

De altfel, aceasta rezultă și în baza teoremei din care reiese că, dacă șirul de variabile

$$x_1, x_2, \ldots, x_n$$

tind în probabilitate, respectiv, către valorile

$$a_1, a_2, \ldots, a_n,$$

atunci polinomul $P(x_1, x_2, \ldots, x_n)$ tinde, de asemenea, în probabilitate către polinomul $P(a_1, a_2, \ldots, a_n)$.

Cu alte cuvinte, în termeni statistici, dacă valorile

$$a_1, a_2, \ldots, a_n$$

sînt, respectiv, estimații corecte ale variabilelor

$$x_1, x_2, \ldots, x_n,$$

atunci polinomul $P(a_1, a_2, \ldots, a_n)$ este o estimație corectă a polinomului $P(x_1, x_2, \ldots, x_n)$.

Se știe că momentele centrate de selecție $\overline{m}_k$ se pot exprima ca niște polinoame de momentele obișnuite de selecție $\overline{M}_k$ $(k=1, 2, \ldots, n)$.

Deoarece am arătat că momentele $\overline{M}_k$ de selecție sînt estimații corecte ale momentelor teoretice, rezultă, în baza celor de mai sus, că și momentele centrate de selecție $\overline{m}_k$ sînt estimații corecte ale momentelor centrate teoretice m_k.

3. Am stabilit că

$$M(\overline{m}_k) = m_k + o\left(\frac{1}{n}\right).$$

Ne-am limitat, însă, numai la o evaluare asimptotică, deoarece după cum am văzut, calculul valorilor medii ale distribuției lui $\overline{m}_k$ este destul de complicat.

Ne propunem, totuși, să determinăm valorile exacte ale lui

$$M(\overline{m}_2),\ M(\overline{m}_3),\ M(\overline{m}_4),$$

adică valoarea medie exactă a primelor momente centrate de selecție, care se întîlnesc mai frecvent în aplicații.

Pornind de la egalitățile

$$\overline{m}_2 = \overline{M}_2 - \overline{M}_1^2,$$

$$\overline{m}_3 = \overline{M}_3 - 3\overline{M}_1\overline{M}_2 + 2\overline{M}_1^3,$$

$$\overline{m}_4 = \overline{M}_4 - 4\overline{M}_1\overline{M}_3 + 6\overline{M}_1^2\overline{M}_2 - 3\overline{M}_1^4,$$

obținem

$$M(\overline{m}_2) = M(\overline{M}_2) - M(\overline{M}_1^2) = m_2 - \frac{m_2}{n} = \frac{n-1}{n} m_2,$$

deoarece am arătat că

$$M(\overline{M}_2) = m_2$$

și

$$M(\overline{M}_1 - M_1^2)^2 = M(X - M_1)^2 = \frac{m_2}{n},$$

presupunînd că

$$M_1 = m = 0.$$

De asemenea, avem

$$M(m_3) = M(\overline{M}_3) - 3M(\overline{M}_1\overline{M}_2) + 2M(\overline{M}_1^3).$$

Insă

$$M(\overline{M}_3) = m_3,$$

$$M(\overline{M}_1\overline{M}_2) = \frac{1}{n^2} M\left(\sum_{i=1}^{n} x_i \sum_{j=1}^{n} x_j\right) = \frac{1}{n^2} \sum_{i=1}^{n} (x_i^3) = \frac{m_3}{n},$$

$$M(\overline{M}_1^3) = \frac{1}{n^3} M\left(\sum_{i=1}^{n} x_i\right)^3 = \frac{1}{n^3} \sum M(x_i^3) = \frac{m^3}{n^2},$$

deoarece restul termenilor obținuți prin ridicarea sumei $\sum_{i=1}^{n} x_i$ la puterea a treia conțin ca factori momentele centrate teoretice de ordinul întîi, care sînt nule.

Prin urmare, deducem

$$M(\overline{m}_3) = m_3 - \frac{3m_3}{n} + \frac{2m_3}{n^2} = \frac{(n-1)(n-2)}{n^2} m_3.$$

In mod analog se obține

$$M(\overline{m}_4) = \frac{(n-1)(n^2-3n+3)}{n^2} m_4 + \frac{3(n-1)(2n-3)}{n^3} m_2^2.$$

Valorile lui $M(\overline{m}_3)$ și $M(\overline{m}_4)$ de mai sus au fost calculate presupunînd, ca și în cazul lui $M(\overline{m}_2)$, că $M_1 = m = 0$, ceea ce înseamnă că $M_k = m_k$, pentru orice k întreg și pozitiv.

Relațiile de mai sus rămîn valabile, evident, și pentru $M_1 = m \neq 0$, deoarece momentele centrate teoretice m_k sînt invariante față de o schimbare a axelor de coordonate.

Observăm din expresiile valorilor medii $M(\overline{m}_2)$, $M(\overline{m}_3)$ și $M(\overline{m}_4)$ că momentele centrate de selecție $\overline{m}_2$, $\overline{m}_3$ și $\overline{m}_4$ nu reprezintă respectiv estimații absolut corecte pentru momentele centrate teoretice m_2, m_3 și m_4. Se pot calcula însă și valorile care reprezintă estimații absolut corecte pentru aceste momente teoretice.

De exemplu, se constată ușor că următoarele expresii reprezintă respectiv, estimații absolut corecte pentru momentele teoretice m_2 și m_3

$$\frac{n}{n-1}\overline{m}_2, \quad \frac{n}{(n-1)(n-2)}\overline{m}_3.$$

§ 6. Momente de selecție pentru o distribuție multidimensională

Să considerăm o distribuție într-un spațiu euclidian cu s dimensiuni și să presupunem că am făcut n selecții independente, care ne-au condus la următoarele rezultate, exprimate prin vectorii

$$P_1(x_{11}, x_{21}, \ldots, x_{s1}),\ P(x_{12}, x_{22}, \ldots, x_{s2}), \ldots, P(x_{1n}, x_{2n}, \ldots, x_{sn}).$$

Toate teoremele pe care le-am demonstrat cu privire la momentele unei distribuții de o variabilă rămîn valabile și pentru momentele unei distribuții multidimensionale. In cele ce urmează, vom demonstra unele dintre aceste teoreme pentru o distribuție cu s dimensiuni.

Teoremă. *Dacă momentele m_i și M_{ij} ale unei distribuții teoretice cu s dimensiuni există, atunci media ei de selecție*

$$\bar{x}_i = \frac{\sum_{j=1}^{n} x_{ij}}{n} \qquad (i = 1, 2, \ldots, s)$$

este o estimație absolut corectă a lui m_i.

Avem evident

$$M(\bar{x}_i) = \frac{1}{n} M\left(\sum_{j=1}^{n} x_{ij}\right) = \frac{1}{n} \sum_{j=1}^{n} M(x_{ij}) = m_i,$$

deoarece, prin definiție, $M(x_{ij}) = m_i$, pentru orice $j = 1, 2, \ldots, n$.

Pe de altă parte,

$$D^2(\bar{x}_i) = M(\bar{x}_i - m_i)^2 = M\left(\frac{\sum_{j=1}^{n} x_{ij}}{n} - m_i\right)^2 = \frac{1}{n^2} M\left[\sum_{j=1}^{n} (x_{ij} - m_i)\right]^2 =$$

$$= \frac{1}{n^2} \sum_{j=1}^{n} M(x_{ij} - m_i)^2 = \frac{m_{ii}}{n},$$

întrucît variabilele $x_{ij} - m_i$ și $x_{hk} - m_h$ fiind independente, momentele $M(x_{ij} - m_i)(x_{hk} - m_h) = M(x_{ij} - m_i)\, M(x_{hk} - m_h) = 0$, iar în baza definiției sale, $M(x_{ij} - m_i)^2 = m_{ii}$, oricare ar fi $i = 1, 2, \ldots, s$).

Din relațiile

$$M(\bar{x}_i) = m_i$$

și

$$\lim_{n\to\infty} D^2(\bar{x}_i)=\lim_{n\to\infty}\frac{m_{ii}}{n}=0,$$

rezultă că media de selecție $\bar{x}_i$ este o estimație absolut corectă a momentului teoretic m_i, pentru orice $i=1, 2, \ldots s$, dacă bineînțeles există m_i și m_{ii} (și, deci, implicit M_{ij}).

Cu aceasta, teorema este demonstrată.

Teoremă. *Dacă există momentele* $M_{r_1r_2\ldots r_s}$ și $M_{2r_12r_2\ldots 2r_s}$ *ale unei distribuții teoretice cu s dimensiuni, atunci momentul de selecție*

$$\overline{M}_{r_1r_2\ldots r_s}=\frac{\sum_{i=1}^{n} x_{1i}^{r_1}x_{2i}^{r_2}\ldots x_{si}^{r_s}}{n}$$

este o estimație absolut corectă a lui $M_{r_1r_2\ldots r_s}$.

In adevăr, avem

$$M(\overline{M}_{r_1r_2\ldots r_s})=\frac{1}{n}M\left(\sum_{i=1}^{n}x_{1i}^{r_1}x_{2i}^{r_2}\ldots x_{s_i}^{r_s}\right)=$$

$$=\frac{1}{n}\sum_{i=1}^{n}M(x_{1i}^{r_1}x_{2i}^{r_2}\ldots x_{si}^{r_s})=M_{r_1r_2\ldots r_s}; \qquad \text{(XII.6,1)}$$

avînd în vedere că, pentru orice $i=1, 2, \ldots, n$, $M(x_{1i}^{r_1}, x_{2i}^{r_2}, \ldots, x_{si}^{r_s})=M_{r_1r_2\ldots r_s}$,

$$D^2(\overline{M}_{r_1r_2\ldots r_s})=M(\overline{M}_{r_1r_2\ldots r_s}-M_{r_1r_2\ldots r_s})^2=$$

$$=M\left(\frac{\sum_{i=1}^{n}x_{1i}^{r_1}x_{2i}^{r_2}\ldots x_{si}^{r_s}}{n}-M_{r_1r_2\ldots r_s}\right)^2=$$

$$=\frac{1}{n^2}M\left[\sum_{i=1}^{n}(x_{1i}^{r_1}x_{2i}^{r_2}\ldots x_{si}^{r_s}-M_{r_1r_2\ldots r_s})\right]=$$

$$=\frac{1}{n^2}\sum_{i=1}^{n}M(x_{1i}^{r_1}x_{2i}^{r_2}\ldots x_{si}^{r_s}-M_{r_1r_2\ldots r_s})^2+$$

$$+\frac{2}{n^2}\sum_{i\neq j}M(x_{1i}^{r_1}x_{2i}^{r_2}\ldots x_{si}^{r_s}-M_{r_1r_2\ldots r_s})(x_{1j}^{r_1}x_{2j}^{r_2}\ldots x_{sj}^{r_s}-M_{r_1r_2\ldots r_s}).$$

Selecțiile fiind independente, urmează că

$$M\left(x_{1i}^{r_1} x_{2i}^{r_2} \dots x_{si}^{r_s} - M_{r_1 r_2 \dots r_s}\right)\left(x_{1j}^{r_1} x_{2j}^{r_2} \dots x_{sj}^{r_s} - M_{r_1 r_2 \dots r_s}\right) =$$

$$= M\left(x_{1i}^{r_1} x_{2i}^{r_2} \dots x_{si}^{r_s} - M_{r_1 r_2 \dots r_s}\right) M\left(x_{1j}^{r_1} x_{2j}^{r_2} \dots x_{sj}^{r_s} - M_{r_1 r_2 \dots r_s}\right) = 0.$$

Prin urmare, obținem

$$D^2\left(\overline{M}_{r_1 r_2 \dots r_s}\right) = \frac{1}{n^2} \sum_{i=1}^{n} M\left(x_{1i}^{r_1} x_{2i}^{r_2} \dots x_{si}^{r_s} - M_{r_1 r_2 \dots r_s}\right)^2 =$$

$$= \frac{1}{n^2} \sum_{i=1}^{n} M\left(x_{1i}^{2r_1} x_{2i}^{2r_2} \dots x_{si}^{2r_s} - 2M_{r_1 r_2 \dots r_s} \cdot x_{1i}^{r_1} x_{2i}^{r_2} \dots x_{si}^{r_s} + M_{r_1 r_2 \dots r_s}^2\right) =$$

$$= \frac{1}{n^2} \sum_{i=1}^{n} \left(M_{2r_1 2r_2 \dots 2r_s} - M_{r_1 r_2 \dots r_s}^2\right) = \frac{M_{2r_1 2r_2 \dots 2r_s} - M_{r_1 r_2 \dots r_s}}{n},$$

de unde rezultă că

$$\lim_{n \to \infty} D^2\left(\overline{M}_{r_1 r_2 \dots r_s}\right) = \lim_{n \to \infty} \frac{M_{2r_1 2r_2 \dots 2r_s} - M_{r_1 r_2 \dots r_s}}{n} = 0. \qquad \text{(XII.6,2)}$$

Atunci, dacă există momentele teoretice $M_{r_1 r_2 \dots r_s}$ și $M_{2r_1 2r_2 \dots 2r_s}$, momentul de selecție $\overline{M}_{r_1 r_2 \dots r_s}$, conform relațiilor (XII.6,1) și (XII.6,2) este o estimație absolut corectă a momentului teoretic $M_{r_1 r_2 \dots r_s}$, ceea ce demonstrează teorema.

Teoremă. *Momentul centrat de selecție*

$$\overline{m}_{r_1 r_2 \dots r_s} = \frac{1}{n} \sum_{i=1}^{n} (x_{1i} - \overline{x}_1)^{r_1} (x_{2i} - \overline{x}_2)^{r_2} \dots (x_{si} - \overline{x}_s)^{r_s}$$

este o estimație corectă a momentului centrat teoretic

$$m_{r_1 r_2 \dots r_s} = M\left(x_{1i} - m_1\right)^{r_1} (x_{2i} - m_2)^{r_2} \dots (x_{si} - m_s)^{r_s},$$

cu condiția ca momentele teoretice care intervin în calcule să existe.

Din teoria calculului probabilităților se cunoaște că, dacă avem un șir de variabile

$$x_1(n), \quad x_2(n), \dots, x_k(n),$$

care depind de parametrul n și pentru care funcția de repartiție

$$F_n(a_1, a_2, \ldots, a_k) \underset{n\to\infty}{\to} F(a_1, a_2, \ldots, a_k),$$

atunci, orice funcție

$$g[x_1(n), x_2(n), \ldots, x_k(n)]$$

măsurabilă Borel tinde, cînd $n\to\infty$, în punctele ei de continuitate în sensul lui Bernoulli, către

$$g(a_1, a_2, \ldots, a_k).$$

Deoarece convergența în sensul lui Bernoulli implică convergența în probabilitate, rezultă că funcția

$$g[x_1(n), x_2(n), \ldots, x_k(n)]$$

tinde, cînd $n\to\infty$, în probabilitate către

$$g(a_1, a_2, \ldots, a_k).$$

Pentru ca această proprietate să fie valabilă, se poate arăta că nu este necesar ca șirul de variabile

$$x_1(n), x_2(n), \ldots, x_k(n)$$

să tindă, cînd $n\to\infty$, în sensul lui Bernoulli către valorile

$$a_1, a_2, \ldots, a_k,$$

ci este suficient numai ca șirul acestor variabile să tindă în probabilitate către șirul valorilor de mai sus.
In termeni statistici, aceasta înseamnă că, dacă șirul de valori

$$a_1, a_2, \ldots, a_k$$

sînt respectiv estimații corecte ale șirului de variabile

$$x_1(n), x_2(n), \ldots, x_k(n),$$

atunci

$$g(a_1, a_2, \ldots, a_k)$$

este, la rîndul ei, o estimație corectă a funcției

$$g[x_1(n), x_2(n), \ldots, x_k(n)].$$

Am demonstrat, că momentele de selecție $\overline{M}_{r_1r_2\ldots r_s}$ sînt estimații corecte ale momentelor teoretice $M_{r_1r_2\ldots r_s}$, dacă există momentele teoretice $M_{r_1r_2\ldots r_s}$ și $M_{2r_1 2r_2\ldots 2r_s}$.

In același timp, se știe că momentele centrate de selecție $\overline{m}_{r_1r_2\ldots r_s}$ se pot exprima ca funcții de momentele necentrate de selecție $\overline{M}_{r_1r_2\ldots r_s}$, de unde deducem, în urma proprietății de mai sus, că momentele centrate de selecție $\overline{m}_{r_1r_2\ldots r_s}$ sînt estimații corecte ale momentelor centrate teoretice, funcții, la rîndul lor, de momentele necentrate teoretice. Teorema este, deci, demonstrată.

Mai mult decît atît, toate noțiunile pe care le-am definit pentru distribuțiile multidimensionale, ca de pildă, coeficientul multiplu de corelație, coeficientul parțial de corelație etc., pot fi estimate în baza unei selecții de mărime n.

Astfel, coeficientul multiplu de corelație este, după cum s-a stabilit la timpul său, definit de expresia

$$R^2_{1.23\ldots s}=1-\frac{|M|}{m_{11}|A|},$$

unde $|M|=|m_{ij}|$ $(i=1, 2, \ldots, s)$ și $|A|=|m_{hk}|$ $(h, k=2, 3\ldots, s)$, m_{ij}, m_{hk} fiind momentele teoretice centrate ale distribuției s-dimensionale.

Atunci, considerînd determinanții

$$|\overline{M}|=|\overline{m}_{ij}| \qquad (i, j=1, 2, \ldots, s)$$

și

$$|\overline{A}|=|\overline{m}_{hk}| \qquad (h, k=2, 3, \ldots, s),$$

unde $\overline{m}_{ij}=\frac{1}{n}\sum_{s=1}^{n}(\overline{x}_{is}-\overline{x}_i)(x_{js}-\overline{x}_j)$, coeficientul de mai jos

$$\overline{R}^2_{1.23\ldots s}=1-\frac{|\overline{M}|}{\overline{m}_{11}|\overline{A}|}$$

este o estimație corectă a coeficientului multiplu de corelație $R^2_{1.23\ldots s}$.

Aceasta rezultă ca o consecință imediată a ultimei teoreme a paragrafului de față, deoarece, după cum se observă, coeficienții $\overline{R}^2_{1\cdot 2\ldots s}$ și $R^2_{1\cdot 2\ldots s}$ sînt, respectiv, funcții de $\overline{m}_{ij}$ și m_{ij}, unde momentele centrate de selecție $\overline{m}_{ij}$ sînt estimații corecte ale momentelor centrate teoretice m_{ij}.

Capitolul XIII

Selecția dintr-o colectivitate normală

§ 1. Considerații generale

Cele mai frecvente cazuri întîlnite în aplicarea metodei selecției corespund cazului cînd distribuția teoretică este o distribuție normală. De exemplu, dacă considerăm erorile de observație la diferitele măsurători, se știe din teoria erorilor că aceste erori sînt distribuite după legea normală. Rezultă astfel un cîmp mare de cercetări, la care selecția dintr-o colectivitate normală se impune în mod natural ca mijloc de investigație.

In teoria generală a selecției am arătat că, pentru o distribuție teoretică normală, putem determina distribuția exactă a unei selecții de mărime n. Așa, de pildă, dacă avem selecțiile independente

$$x_1,\ x_2, \ldots, x_n$$

și distribuția teoretică este o distribuție normală $N(m, \sigma)$, adică admite funcția de frecvență

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-\frac{(x-m)^2}{2\sigma^2}},$$

atunci media de selecție

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

are distribuția normală $N\left(m, \frac{\sigma}{\sqrt{n}}\right)$.

De asemenea, din studiul distribuțiilor înrudite cu distribuția lui Helmert-Pearson rezultă că momentul de selecție de ordinul al doilea

$$\overline{M}_2 = \frac{x_1^2+x_2^2+\ldots+x_n^2}{n}$$

dă naștere unei distribuții cu funcția de frecvență

$$h(x) = \frac{n^{\frac{n}{2}}}{2^{\frac{n}{2}}\sigma^n \Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1} \cdot e^{-\frac{nx}{2\sigma^2}} \qquad (x \geqslant 0).$$

In cele ce urmează, vom considera și alte funcții de selecție și vom arăta unele proprietăți în legătură cu ele.

§ 2. Distribuția momentului centrat de selecție de ordinul al doilea

1. In ipoteza că distribuția teoretică urmează o lege normală $N(0, \sigma)$, ne propunem să determinăm distribuția momentului centrat de selecție de ordinul al doilea

$$\overline{m}_2 = \frac{\sum_{k=1}^{n} (x_k-\overline{x})^2}{n}.$$

Putem introduce și notația

$$\overline{\sigma}^2 = \frac{\sum_{k=1}^{n} (x_k-\overline{x})^2}{n}$$

pentru a arăta că $\overline{\sigma}^2$ este dispersia corespunzătoare distribuției empirice cu masele $\frac{1}{n}$ așezate, respectiv, în punctele x_k $(k=1, 2, \ldots, n)$. Avem identitățile cunoscute

$$\sum_{k=1}^{n} (x_k-\overline{x})^2 = \sum_{k=1}^{n} x_k^2 - \frac{1}{n}\left(\sum_{k=1}^{n} x_k\right)^2$$

și

$$\sum_{k=1}^{n}(x_k-\bar{x})^2=\frac{n-1}{n}\left(x_1-\frac{\sum_{k=2}^{n}x_k}{n-1}\right)^2+$$

$$+\frac{n-2}{n-1}\left(x_2-\frac{\sum_{k=3}^{n}x_k}{n-2}\right)^2+\ldots+\frac{1}{2}(x_{n-1}-x_n)^2.$$

De unde rezultă

$$\sum_{k=1}^{n}x_k^2=\frac{1}{n}\left(\sum_{k=1}^{n}x_k\right)^2+\frac{n-1}{n}\left(x_1-\frac{\sum_{k=2}^{n}x_k}{n-1}\right)^2+$$

$$+\frac{n-2}{n-1}\left(x_2-\frac{\sum_{k=3}^{n}x_k}{n-2}\right)^2+\ldots+\frac{1}{2}(x_{n-1}-x_n)^2.$$

Obervăm că membrul al doilea al acestei egalități este o sumă de n forme pătratice pozitiv definite, fiecare avînd rangul egal cu unu. Deoarece suma rangurilor acestor forme pătratice pozitiv definite este egală cu n, iar variabilele x_k $(k=1, 2, \ldots, n)$ urmează fiecare o lege normală $N(0, \sigma)$, atunci după teorema lui Cochran, variabilele

$$y_1=\sqrt{\frac{n-1}{n}}\left(x_1-\frac{\sum_{k=2}^{n}x_k}{n-1}\right),$$

$$y_2=\sqrt{\frac{n-2}{n-1}}\left(x_2-\frac{\sum_{k=3}^{n}x_k}{n-2}\right),$$

$$\ldots\ldots\ldots\ldots$$

$$y_{n-1}=\frac{1}{\sqrt{2}}(x_{n-1}-x_n),$$

$$y_n=\frac{1}{\sqrt{n}}\sum_{k=1}^{n}x_k$$

sînt independente între ele.

Pe de altă parte, avem

$$\overline{\sigma}^2 = \frac{y_1^2+y_2^2+\ldots+y_{n-1}^2}{n},$$

iar

$$M(y_1) = \sqrt{\frac{n-1}{n}}\left[M(x_1) - \frac{1}{n-1}\sum_{k=2}^{n} M(x_k)\right] = 0,$$

fiindcă $M(x_k) = m$, pentru orice k.
In mod analog se demonstrează că

$$M(y_i) = 0 \qquad (i = 1, 2, .., , n-1).$$

Notînd cu σ_1^2 dispersia variabilei y_1, deducem

$$\sigma_1^2 = M\left[\sqrt{\frac{n-1}{n}}\left(x_1 - \frac{\sum_{k=2}^{n} x_k}{n-1}\right)\right]^2 = \frac{n-1}{n} M\left(x_1 - \frac{\sum_{k=2}^{n} x_k}{n-1}\right)^2 =$$

$$= \frac{n-1}{n} M\left[x_1^2 + \frac{\sum_{k=2}^{n} x_k^2}{(n-1)^2}\right] = \frac{n-1}{n}\sigma^2\left(1 + \frac{1}{n-1}\right) = \sigma^2,$$

ținînd seama că variabilele $x_k\,(k = 1, 2, \ldots, n)$ sînt independente între ele și fiecare are o distribuție normală $N(0, \sigma)$.
La fel se arată că

$$D^2(y) = \sigma^2 \qquad (i = 1, 2, 3, \ldots, n-1).$$

Am demonstrat, astfel, că variabilele

$$y_1, y_2, \ldots, y_{n-1}$$

sînt independente între ele și fiecare dă naștere unei distribuții normale $N(0, \sigma)$. De aici urmează că variabila aleatorie

$$\overline{\sigma} = \frac{\sqrt{y_1^2+y_2^2+\ldots+y_{n-1}^2}}{\sqrt{n}},$$

în baza celor stabilite la studiul distribuțiilor înrudite cu distribuția Helmert-Pearson (cap, XI, § 8), admite distribuția a cărei funcție de frecvență este

$$g(x)=\frac{2n^{\frac{n-1}{2}}}{2^{\frac{n-1}{2}}\sigma^{n-1}\Gamma\left(\frac{n-1}{2}\right)}\cdot x^{n-2}\cdot e^{-\frac{nx^2}{2\sigma^2}} \qquad (x\geqq 0).$$

In ceea ce privește variabila aleatorie $n\sigma^2=y_1^2+y_2^2+\ldots+y_{n-1}^2$, ea urmează legea $H(n-1, \sigma)$ adică o lege Helmert-Pearson cu $n-1$ grade de libertate.
Mai mult decît atît, din faptul că variabilele $y_i\,(i=1, 2, \ldots, n)$ sînt independente între ele, reiese că, la rîndul lor, variabilele

$$\bar{\sigma}^2=\frac{y_1^2+y_2^2+\ldots+y_{n-1}^2}{n}$$

și

$$\bar{x}=\frac{y_n}{\sqrt{n}}$$

admit distribuții independente.
Prin urmare, putem enunța teorema următoare.

Teoremă. *Dacă*

$$x_1,\ x_2, \ldots, x_n$$

sînt selecții independente dintr-o colectivitate normală $N(0, \sigma)$, *atunci:*

a) *media de selecție* $\bar{x}$ *admite distribuția normală* $N\left(0, \frac{\sigma}{\sqrt{n}}\right)$;

b) *variabila* $\bar{\sigma}=\sqrt{m_2}$ *are o distribuție a cărei funcție de frecvență este*

$$g(x)=\frac{2n^{\frac{n-1}{2}}}{2^{\frac{n-1}{2}}\sigma^{n-1}\Gamma\left(\frac{n-1}{2}\right)}\cdot x^{n-2}\cdot e^{-\frac{nx^2}{2\sigma^2}} \qquad (x\geqq 0);$$

c) *variabilele aleatorii* $\bar{x}$ *și* $\bar{m}_2$ *sînt variabile independente.*

Trebuie să reținem faptul important că, pentru o selecție de mărime n dintr-o colectivitate $N(0, \sigma)$, momentul de selecție de ordinul

al doilea urmează o lege de tipul Helmert-Pearson cu n grade de libertate, pe cînd momentul centrat de selecție de ordinul al doilea urmează o lege de tipul Helmert-Pearson cu $n-1$ grade de libertate.

Am calculat mai înainte covarianța pentru două momente centrate de selecție oarecare $\overline{m}_h$ și $\overline{m}_k$. In cazul particular cînd $h=1$ și $k=2$, obținem

$$M[(\overline{x}-M(\overline{x}))(\overline{m}_2-M(\overline{m}_2))]=\frac{n-1}{n}m_3.$$

Dacă distribuția teoretică este simetrică, atunci, evident, $m_3=0$ și, deci, variabilele $\overline{x}$ și $\overline{m}_2$ sînt necorelate. Observăm astfel și pe această cale, independența variabilelor $\overline{x}$ și $\overline{m}_2$ pentru o distribuție teoretică normală.

2. Vom da acum o altă demonstrație pentru determinarea distribuției variabilei $\overline{x}$ și $\sqrt{\overline{m}_2}$.

Fie

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}\cdot e^{-\frac{(x-m)^2}{2\sigma^2}}$$

funcția de frecvență a distribuției teoretice, adică a unei distribuții normale $N(m, \sigma)$.

Cele n selecții independente

$$x_1, x_2, \ldots, x_n$$

le considerăm coordonatele unui punct $P(x_1, x_2, \ldots, x_n)$ dintr-un spațiu euclidian cu n dimensiuni. Variabilele $x_i (i=1, 2, \ldots, n)$ fiind independente, masa concentrată în punctul $P(x_1, x_2, \ldots, x_n)$ este

$$\frac{1}{(2\pi)^{\frac{n}{2}}\sigma^n}\cdot e^{-\frac{\sum_{k=1}^{n}(x_k-m)^2}{2\sigma^2}}\cdot dx_1\, dx_2 \ldots dx_n$$

sau, cu alte cuvinte, probabilitatea ca, făcînd n selecții independente, să obținem valori cuprinse între x_i și $x_i+dx_i (i=1, 2, \ldots, n)$.

In baza identității

$$\sum_{k=1}^{n}(\bar{x}_k-m)^2=n(\bar{x}-m)^2+n\bar{\sigma}^2,$$

probabilitatea elementară de mai sus devine

$$\frac{1}{(2\pi)^{\frac{n}{2}}\sigma^n}\cdot e^{-\frac{n(\bar{x}-m)^2}{2\sigma^2}}\cdot e^{-\frac{n\bar{\sigma}^2}{2\sigma^2}}\cdot dx_1\,dx_2\ldots dx_n.$$

Să introducem notația

$$x_k=\bar{x}+\sqrt{n}\cdot\bar{\sigma}\xi_k \qquad (k=1,\ 2,\ \ldots,\ n).$$

Avînd în vedere expresia lui $\bar{x}$ și $\bar{\sigma}$ în raport cu x_k, rezultă

$$\sum_{k=1}^{n}\xi_k=0 \qquad \text{(XIII.2,1)}$$

și

$$\sum_{k=1}^{n}\xi_k^2=1 \qquad \text{(XIII.2,2)}$$

Să efectuăm în spațiul n-dimensional considerat o schimbare de variabile, astfel ca, în locul variabilelor

$$x_1,\ x_2,\ \ldots,\ x_n,$$

să avem variabilele

$$\xi_1,\ \xi_2,\ \ldots,\ \xi_{n-2},\ \bar{x},\ \bar{\sigma}.$$

In virtutea relațiilor (XIII.2,1) urmează că variabilele ξ_{n-1} și ξ_n sînt funcții de variabilele $\xi_1, \xi_2, \ldots, \xi_{n-2}$. Atunci, derivînd relațiile (XIII.2,1) și (XIII.2,2), deducem

$$1+\frac{\partial\xi_{n-1}}{\partial\xi_k}+\frac{\partial\xi_n}{\partial\xi_k}=0 \qquad (k=1,\ 2,\ \ldots,\ n) \qquad \text{(XIII.2,3)}$$

și

$$\xi_k+\xi_{n-1}\cdot\frac{\partial\xi_{n-1}}{\partial\xi_k}+\xi_n\cdot\frac{\partial\xi_n}{\partial\xi_k}=0 \qquad (k=1,\ 2,\ \ldots,\ n). \qquad \text{(XIII.2,4)}$$

Pentru efectuarea transformării, trebuie să calculăm determinantul funcțional $\frac{D(x_1, x_2, \dots, x_n)}{D(\xi_1, \xi_2, \dots, \xi_{n-2}, \bar{x}, \bar{\sigma})}$. Obținem astfel

$$\frac{D(x_1, x_2, \dots, x_n)}{D(\xi_1, \xi_2, \dots, \xi_{n-2}, \bar{x}, \bar{\sigma})} =$$

$$= n^{\frac{n}{2}} \begin{vmatrix} \sqrt{n}\bar{\sigma} & 0 & 0 & \dots & 0 & \xi_1 & 1 \\ 0 & \sqrt{n}\bar{\sigma} & 0 & \dots & 0 & \xi_2 & 1 \\ 0 & 0 & \sqrt{n}\bar{\sigma} & \dots & 0 & \xi_3 & 1 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \sqrt{n}\bar{\sigma} & \xi_{n-2} & 1 \\ \sqrt{n}\bar{\sigma}\frac{\partial \xi_{n-1}}{\partial \xi_1} & \sqrt{n}\bar{\sigma}\frac{\partial \xi_{n-1}}{\partial \xi_2} & \sqrt{n}\bar{\sigma}\frac{\partial \xi_{n-1}}{\partial \xi_3} & \dots & \sqrt{n}\bar{\sigma}\frac{\partial \xi_{n-1}}{\partial \xi_{n-2}} & \xi_{n-1} & 1 \\ \sqrt{n}\bar{\sigma}\frac{\partial \xi_n}{\partial \xi_1} & \sqrt{n}\bar{\sigma}\frac{\partial \xi_n}{\partial \xi_2} & \sqrt{n}\bar{\sigma}\frac{\partial \xi_n}{\partial \xi_3} & \dots & \sqrt{n}\bar{\sigma}\frac{\partial \xi_n}{\partial \xi_{n-2}} & \xi_n & 1 \end{vmatrix}.$$

Adunînd elementele tuturor liniilor la ultima linie și ținînd seama de relațiile (XIII.2,1) și (XIII.2,3) rezultă

$$\frac{D(x_1, x_2, \dots, x_n)}{D(\xi_1, \xi_2, \dots, \xi_{n-2}, \bar{x}, \bar{\sigma})} =$$

$$= -n^{\frac{n}{2}+1} \begin{vmatrix} \sqrt{n}\bar{\sigma} & 0 & 0 & \dots & 0 & \xi_1 \\ 0 & \sqrt{n}\bar{\sigma} & 0 & \dots & 0 & \xi_2 \\ 0 & 0 & \sqrt{n}\bar{\sigma} & \dots & 0 & \xi_3 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \sqrt{n}\bar{\sigma}\frac{\partial \xi_{n-1}}{\partial \xi_1} & \sqrt{n}\bar{\sigma}\frac{\partial \xi_{n-1}}{\partial \xi_2} & \sqrt{n}\bar{\sigma}\frac{\partial \xi_{n-1}}{\partial \xi_3} & \dots & \sqrt{n}\bar{\sigma}\frac{\partial \xi_{n-1}}{\partial \xi_{n-2}} & \xi_{n-1} \end{vmatrix}.$$

Dacă din ultima coloană înmulţită cu $\sqrt{n}\,\bar{\sigma}$ scădem primele $n-1$ coloane înmulţite respectiv cu $\xi_1, \xi_2, \ldots, \xi_{n-2}$, avem

$$\frac{D(x_1, x_2, \ldots, x_n)}{D(\xi_1, \xi_2, \ldots, \xi_{n-2}, \bar{x}, \bar{\sigma})} = n^{n-\frac{1}{2}}\bar{\sigma}^{n-3}A,$$

unde am notat

$$A = n^{\frac{1}{2}}\bar{\sigma}\left(\xi_{n-1} - \xi_1\frac{\partial\xi_{n-1}}{\partial\xi_1} - \xi_2\frac{\partial\xi_{n-1}}{\partial\xi_2} - \ldots - \xi_{n-2}\frac{\partial\xi_{n-1}}{\partial\xi_{n-2}}\right).$$

Insă, din relaţiile (XIII.2,3) şi (XIII.2,4), deducem

$$-\xi_k\frac{\partial\xi_{n-1}}{\partial\xi_k} = \xi_k\cdot\frac{\xi_k-\xi_n}{\xi_{n-1}-\xi_n}.$$

Inlocuind valorile lui $-\xi_k\cdot\frac{\partial\xi_{n-1}}{\partial\xi_k}$ $(k=1, 2, \ldots, n-2)$, expresia lui A devine

$$A = n^{\frac{1}{2}}\bar{\sigma}\frac{\xi_{n-1}(\xi_{n-1}-\xi_n)+\ldots+\xi_1(\xi_1-\xi_n)}{\xi_{n-1}-\xi_n} = n^{\frac{1}{2}}\bar{\sigma}\frac{1}{\xi_{n-1}-\xi_n}$$

şi, deci, determinantul funcţional este egal cu

$$\frac{D(x_1, x_2, \ldots, x_n)}{D(\xi_1, \xi_2, \ldots, \xi_{n-2}, \bar{x}, \bar{\sigma})} = n^n\bar{\sigma}^{n-2}\frac{1}{|\xi_{n-1}-\xi_n|},$$

deoarece funcţia de frecvenţă este pozitivă.

Observăm că fiecărui punct $(x_1, x_2, \ldots, x_n)$ îi corespunde un singur punct $(\xi_1, \xi_2, \ldots, \xi_{n-2}, \bar{x}, \bar{\sigma})$ şi fiecărui punct $(\xi_1, \xi_2, \ldots, \xi_{n-2}, \bar{x}, \bar{\sigma})$ îi corespund două puncte $(x_1, x_2, \ldots, x_n)$ în care, însă, $|\xi_{n-1}-\xi_n|$ păstrează aceeaşi valoare.

Rezultă, atunci, că probabilitatea elementară

$$\frac{1}{(2\pi)^{\frac{n}{2}}\sigma^n}\cdot e^{-\frac{n(\bar{x}-m)}{2\sigma^2}}\cdot e^{-\frac{n\bar{\sigma}^2}{2\sigma^2}}\cdot dx_1\,dx_2\ldots dx_n,$$

prin transformarea efectuată se scrie

$$\frac{2n^n\,\bar{\sigma}^{n-2}}{(2\pi)^{\frac{n}{2}}\,\sigma^n}\cdot\frac{1}{|\xi_{n-1}-\xi_n|}e^{-\frac{n(\bar{x}-m)^2}{2\sigma^2}}\cdot e^{-\frac{n\bar{\sigma}^2}{2\sigma^2}}\,d\bar{x}\,d\bar{\sigma}\,d\xi_1\ldots d\xi_{n-2}=$$

$$=\frac{n^{\frac{1}{2}}}{(2\pi)^{\frac{1}{2}}}e^{-\frac{n(\bar{x}-m)^2}{2\sigma^2}}\cdot d\bar{x}\cdot\frac{2n^{\frac{n-1}{2}}}{2^{\frac{n-1}{2}}\cdot\sigma^{n-1}\,\Gamma\left(\frac{n-1}{2}\right)}\cdot\bar{\sigma}^{n-2}\times$$

$$\times e^{-\frac{n\bar{\sigma}^2}{2\sigma^2}}\,d\bar{\sigma}\,\frac{n^{\frac{n}{2}}\,\Gamma\left(\frac{n-1}{2}\right)}{\pi^{\frac{n-1}{2}}}\cdot\frac{1}{|\xi_{n-1}-\xi_n|}\cdot d\xi_1\,d\xi_2\ldots d\xi_{n-2}.$$

Cum însă

$$\frac{n^{\frac{1}{2}}}{(2\pi)^{\frac{1}{2}}\,\sigma}\int_{-\infty}^{\infty}e^{-\frac{n(\bar{x}-m)^2}{2\sigma^2}}\,d\bar{x}=1$$

și

$$\frac{2n^{\frac{n-1}{2}}}{2^{\frac{n-1}{2}}\,\sigma^{n-1}\,\Gamma\left(\frac{n-1}{2}\right)}\int_0^{\infty}\bar{\sigma}^{n-2}e^{-\frac{n\bar{\sigma}^2}{2\sigma^2}}\,d\bar{\sigma}=1,$$

urmează că funcțiile

$$f(x)=\frac{n^{\frac{1}{2}}}{\sigma\,(2\pi)^{\frac{n}{2}}}e^{-\frac{n(\bar{x}-m)^2}{2\sigma^2}}$$

și

$$g(x)=\frac{2n^{\frac{n-1}{2}}}{2^{\frac{n-1}{2}}\,\sigma^{n-1}\,\Gamma\left(\frac{n-1}{2}\right)}\bar{\sigma}^{n-2}\cdot e^{-\frac{n\bar{\sigma}^2}{2\sigma^2}}$$

reprezintă funcții de frecvență corespunzătoare distribuțiilor variabilelor $\bar{x}$, respectiv $\bar{\sigma}$.

După cum se vede, prima este funcția de frecvență a distribuției normale $N\left(m, \frac{\sigma}{\sqrt{n}}\right)$, iar a doua este funcția de frecvență a unei distribuții înrudite cu distribuția Helmert-Pearson cu $n-1$ grade de libertate, adică am regăsit funcțiile de frecvență ale variabilelor $\bar{x}$ și $\sqrt{\overline{m}_2}$ stabilite prin teorema de la punctul 1 al acestui paragraf.

§ 3. Distribuția raportului dintre abaterea mediei de selecție și valoarea medie pătratică de selecție

Am arătat (cap. XI, § 9) că, dacă X și Y sînt două variabile aleatorii independente, care urmează fiecare respectiv legile $N(0, \sigma)$ și $H(s, \sigma)$, atunci variabila

$$t = \frac{X}{\sqrt{\frac{\overline{Y}}{s}}}$$

urmează distribuția lui Student cu cinci grade de libertate, a cărei funcție de frecvență este

$$g(t) = \frac{\Gamma\left(\frac{s+1}{2}\right)}{\sqrt{\pi s}\,\Gamma\left(\frac{s}{2}\right)}\left(1+\frac{t^2}{s}\right)^{-\frac{s+1}{2}}.$$

Folosind acest rezultat, se demonstrează imediat teorema următoare.

Teoremă. *Dacă avem selecțiile independente*

$$x_1, x_2, \ldots, x_n,$$

luate dintr-o distribuție teoretică normală $N(m, \sigma)$, *raportul*

$$t = \sqrt{n-1} \cdot \frac{\bar{x}-m}{\sqrt{\overline{m}_2}}$$

urmează distribuția lui Student cu $n-1$ *grade de libertate, avînd funcția de frecvență*

$$g(t) = \frac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{\pi(n-1)}\,\Gamma\left(\frac{n-1}{2}\right)}\left(1+\frac{t^2}{n-1}\right)^{-\frac{n}{2}}. \qquad \text{(XIII.3,1)}$$

In adevăr, putem scrie

$$t=\frac{\sqrt{n}\,(\bar{x}-m)}{\sqrt{\frac{n}{n-1}\,\overline{m}_2}}\,.$$

Dacă notăm

$$X=\sqrt{n}\,(\bar{x}-m),\quad Y=\sum_{i=1}^{n}(x_i-\bar{x})=n\,\overline{m}_2\,,$$

atunci raportul de mai sus devine

$$t=\frac{X}{\sqrt{\frac{Y}{n-1}}}\,.$$

Se constată uşor că avem

$$M(X)=0,\quad D^2(X)=\sigma^2=m_2\,,$$

de unde rezultă, pe de o parte, că variabila aleatorie X este distribuită după legea normală $N(0,\sigma)$, iar pe de altă parte, după teorema din paragraful precedent, variabila aleatorie Y urmează, evident, o lege Helmert-Pearson cu $n-1$ grade de libertate.

Pe de altă parte, am demonstrat că $\bar{x}$ şi $\overline{m}_2$ sînt variabile independente între ele, X şi Y sînt deci şi ele independente între ele.

Rezultă astfel că raportul considerat are o distribuţie Student cu $n-1$ grade de libertate a cărei funcţie de frecvenţă este dată de formula (XIII.3,1).

La cap. XI, § 11, am dat acestei teoreme o altă demonstraţie în care nu am ţinut seamă de faptul că variabilele $\bar{x}$ şi $\overline{m}_2$ sînt independente.

§ 4. Cazul unei distribuţii normale multidimensionale

1. Proprietăţile pe care le-am arătat pentru selecţiile dintr-o colectivitate normală cu o dimensiune se păstrează şi pentru cazul cînd distribuţia teoretică este o distribuţie normală cu mai multe dimensiuni.

Trecerea de la o singură dimensiune la mai multe dimensiuni nu este însă lipsită de dificultăţi şi conduce la rezultate noi, importante în special pentru aplicaţii.

Teoremă. *Dacă variabilele independente* $x_1, x_2, \ldots, x_n$ *sînt distribuite după legea normală*

$$\frac{|A|^{\frac{1}{2}}}{(2\pi)^{\frac{s}{2}}} \cdot e^{-\frac{1}{2}x'Ax},$$

unde $A = \| a_{ij} \| (i, j = 1, 2, \ldots, s)$ *este o formă pătratică definită pozitivă, variabila*

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

urmează legea normală cu funcţia de frecvenţă

$$\frac{n^{\frac{s}{2}} |A|^{\frac{1}{2}}}{(2\pi)^{\frac{s}{2}}} e^{-\frac{1}{2}x'nAx}.$$

Să însemnăm prin $\varphi(t)$ funcţia caracteristică a variabilei $\bar{x}$. Avem

$$\varphi(t) = M\left(e^{it' \frac{1}{n} \sum_{j=1}^{n} x_j}\right) = \left[M\left(e^{it' \frac{x_j}{n}}\right)\right]^n.$$

Am arătat însă la studiul distribuţiei normale că

$$M\left(e^{it'x_j}\right) = e^{-\frac{1}{2}t'A^{-1}t}.$$

Rezultă:

$$M\left(e^{it' \frac{x_j}{n}}\right) = e^{-\frac{1}{2n^2}t'A^{-1}t},$$

de unde

$$\varphi(t) = e^{-\frac{1}{2n}t'A^{-1}t},$$

adică tocmai funcţia caracteristică a distribuţiei normale cu funcţia de frecvenţă

$$\frac{n^{\frac{s}{2}}\,|A|^{\frac{1}{2}}}{(2\pi)^{\frac{s}{2}}}\cdot e^{-\frac{1}{2}x'nAx}.$$

Am presupus în enunţul teoremei că vectorul mediu este nul. După cum ştim, aceasta nu înseamnă o restrîngere. După cum rezultă din demonstraţie, dacă distribuţia teoretică este

$$\frac{|A|^{\frac{1}{2}}}{(2\pi)^{\frac{s}{2}}}\cdot e^{-\frac{1}{2}(x-m)'A(x-m)},$$

distribuţia mediei de selecţie $\bar{x}$ este

$$\frac{n^{\frac{s}{2}}\,|A|^{\frac{1}{2}}}{(2\pi)^{\frac{s}{2}}}\cdot e^{-\frac{1}{2}(x-m)'nA(x-m)}.$$

Media selecţiei este, prin urmare, egală cu media distribuţiei teoretice. In schimb, momentele centrate de ordinul al doilea ale selecţiei sînt egale cu momentele centrate respective ale distribuţiei teoretice înmulţite cu $\frac{1}{n}$.

2. Pentru cazul unidimensional, am arătat care este distribuţia momentului centrat de selecţie de ordinul al doilea, precum şi faptul important că media de selecţie este independentă de momentul centrat de ordinul al doilea de selecţie.

Vom generaliza aceste rezultate în spaţiul euclidian cu s dimensiuni. Fie

$$f(x)=\frac{|A|^{\frac{1}{2}}}{(2\pi)^{\frac{s}{2}}}e^{-\frac{1}{2}x'Ax}$$

funcţia de frecvenţă a unei distribuţii normale de rangul s în R_s. Efectuăm o selecţie formată din n variabile independente

$$x_1, x_2, \ldots, x_n,$$

unde

$$x_k=(x_{1k}, x_{2k}, \ldots, x_{sk}) \qquad (k=1,2,\ldots,n).$$

Vectorul mediu $\bar{x}$

$$\bar{x}=\frac{x_1+x_2+\ldots+x_n}{n}$$

are deci componentele

$$\bar{x}_j=\frac{1}{n}\sum_{i=1}^{n} x_{ij} \qquad (j=1,2,\ldots,s).$$

Momentele centrate de selecție de ordinul al doilea sînt

$$y_{ij}=\frac{1}{2}\sum_{h=1}^{n}(x_{ih}-\bar{x}_i)(x_{jh}-\bar{x}_j) \qquad (i,j=1,2,\ldots,s).$$

Ne propunem să determinăm distribuția mixtă a variabilelor $\bar{x}_j$, y_{ij} $(i=1,2,\ldots,s;\ j=1,2,\ldots,s)$. Avem în total s variabile $\bar{x}_j$ și $\frac{s(s-1)}{2}$ variabile y_{ij}. Distribuția de care ne ocupăm este deci cu $\frac{s(s+1)}{2}$ dimensiuni.

Vom demonstra în prealabil următoarea lemă.

Lemă. *Dacă:*

a) *variabilele* x_{ij} $(i,j=1,2,\ldots,s)$, $(x_{ij}=x_{ji})$ *sînt distribuite după legea lui Wishart cu* n *grade de libertate cu funcția de frecvență*

$$\frac{|A|^{\frac{n}{2}}}{A_{\frac{n}{2}}}|X|^{\frac{n-s-1}{2}}\cdot e^{-\frac{1}{2}\sum_{i,j=1}^{s} a_{ij}x_{ij}},$$

unde

$$X=\|x_{ij}\|, \quad A=\|a_{ij}\| \qquad (i,j=1,2,\ldots,s),$$

$$A_p=\pi^{\frac{s(s-1)}{4}}\cdot 2^{ps}\,\Gamma(p)\,\Gamma\left(p-\frac{1}{2}\right)\ldots\Gamma\left(p-\frac{s-1}{2}\right);$$

b) *variabilele* y_i $(i=1,2,\ldots,s)$ *sînt distribuite normal cu funcția de frecvență*

$$\frac{|A|^{\frac{1}{2}}}{(2\pi)^{\frac{s}{2}}} e^{-\frac{1}{2}y'Ay};$$

c) *variabilele* x_{ij} $(i,j=1,2,\ldots,s)$ *sînt independente de variabilele* y_i $(i=1,2,\ldots,s)$,
variabila

$$T^2 = n \sum_{i,j=1}^{s} \frac{|X_{ij}|}{|X|} y_i y_j,$$

unde $|X_{ij}|$ *este minorul elementului* x_{ij} *din determinantul* $|X|$, *urmează legea lui Hotelling cu n grade de libertate.*

Din teorema lui Rao (cap. XI, § 18) rezultă că, dacă $x_1, x_2, \ldots, x_n$ sînt n variabile normale independente, avînd funcția de frecvență

$$\frac{|A|^{\frac{1}{2}}}{(2\pi)^{\frac{s}{2}}} e^{-\frac{1}{2}x'Ax}$$

și însemnînd prin Y_p matricea

$$Y_p = \sum_{\alpha=1}^{p} \| x'_\alpha x_\alpha \|,$$

unde

$$x_k = \left\| \begin{matrix} x_{1k} \\ x_{2k} \\ \vdots \\ x_{sk} \end{matrix} \right\|,$$

variabila T_p, definită prin egalitatea

$$\frac{|Y_p|}{|Y_{p+1}|} = \frac{1}{1+\frac{T_p^2}{p}},$$

urmează legea lui Hotelling cu p grade de libertate.

Un calcul simplu de identificare ne arată că

$$|Y_{p+1}| = |Y_p| + \sum_{i,j} |Y_{p,ij}| x_{i,p+1} x_{j,p+1},$$

unde $|Y_{p,ij}|$ este minorul elementului din linia i şi coloana j din determinantul $|Y_p|$.
Rezultă astfel

$$T_p^2 = p \cdot \sum_{i,j} \frac{|Y_{p,ij}|}{|Y_p|} x_{i,p+1} x_{j,p+1}.$$

Cum variabilele x_i sînt independente, nu avem decît să punem $p=n$ şi să însemnăm

$$x_{i,p+1} = y_i \qquad (i=1, 2, \ldots, s)$$

pentru a obţine lema enunţată.
Prin definiţie, funcţia caracteristică a variabilelor $\bar{x}_j (j=1, 2, \ldots, s)$ şi a variabilelor $y_{hk} (h, k=1, 2, \ldots, s)$ este

$$\varphi(t) = \frac{|A|^{\frac{n}{2}}}{(2\pi)^{\frac{ns}{2}}} \int\int \ldots \int_{R_s} e^{i \sum_{j=1}^{s} t_j \bar{x}_j + i \sum_{h,k=1}^{s} \delta_{hk} t_{hk} y_{hk} - \frac{1}{2} \sum_{h=1}^{n} \sum_{i,j=1}^{s} a_{ij} x_{ih} x_{jh}} \prod_{i,j=1}^{s} dx_{ij},$$

unde

$$\delta_{hk} = 1, \quad \text{pentru} \quad h = k,$$

$$\delta_{hk} = \frac{1}{2}, \quad \text{pentru} \quad h \neq k.$$

In spaţiul R_m să considerăm punctele

$$M_h (x_{h1}, x_{h2}, \ldots, x_{hn}) \qquad (h=1, 2, \ldots, s).$$

Dacă efectuăm în R_n o transformare ortogonală, aceste puncte devin respectiv

$$M'_h (v_{h1}, v_{h2}, \ldots, v_{hn}) \qquad (h=1, 2, \ldots, s).$$

Deoarece distanţele de la origine la punctele M_h rămîn prin această transformare invariante, precum şi unghiurile pe care le fac două direcţii OM_h şi OM_k, rezultă că avem relaţiile

$$\sum_{h=1}^{n} x_{ih} x_{jh} = \sum_{h=1}^{n} v_{ih} v_{jh} \qquad (i, j=1, 2, \ldots, s). \qquad \text{(XIII.4,1)}$$

Să mai observăm că

$$y_{ij}=\frac{1}{n}\sum_{h=1}^{n}x_{ih}x_{jh}-\bar{x}_i\bar{x}_j.$$

Să ne alegem transformarea ortogonală în așa fel încît să avem

$$v_{j1}=\sqrt{n}\cdot\bar{x}_j \qquad (j=1,\ 2,\ \ldots,\ s).$$

Vom avea atunci, în baza relațiilor precedente,

$$y_{ij}=\frac{1}{n}\sum_{h=1}^{n}v_{ih}v_{jh}-\frac{v_{i1}v_{j1}}{n}=\frac{1}{n}\sum_{h=2}^{n}v_{ij}v_{jh}.$$

Astfel că, în baza transformării ortogonale efectuate funcția caracteristică devine

$$\varphi(t)=\varphi_1(t)\,\varphi_2(t), \tag{XIII.4,2}$$

unde

$$\varphi_1(t)=\frac{|A|^{\frac{1}{2}}}{(2\pi)^{\frac{s}{2}}}\iint\ldots\int_{R_s}e^{\frac{1}{\sqrt{n}}\sum_{j=1}^{s}t_jv_j-\frac{1}{2}\sum_{i,j=1}^{s}a_{ij}v_{i1}v_{j1}}\prod_{i=1}^{s}dv_{i1},$$

$$\varphi_2(t)=$$

$$=\frac{|A|^{\frac{n-1}{2}}}{(2\pi)^{\frac{(n-1)s}{2}}}\iint\ldots\int_{R_s}e^{\frac{i}{n}\sum_{h,k=1}^{s}\delta_{hk}t_{hk}\sum_{j=2}^{s}v_{hj}v_{kj}-\frac{1}{2}\sum_{i,j}a_{ij}\sum_{h=2}^{n}v_{ih}v_{jh}}\prod dv_{ij}$$

$$(i=1,\ 2,\ \ldots,\ s)\quad (j=2,\ 3,\ \ldots,\ n).$$

$\varphi_1(t)$ conține numai variabilele t_j $(j=1,\ 2,\ \ldots,\ s)$. Ea este deci funcția caracteristică referitoare la variabilele $\bar{x}_j$ $(j=1,\ 2,\ \ldots,\ s)$. $\varphi_2(t)$ este o funcție numai de variabilele t_{ij}. Ea este deci funcția caracteristică a variabilelor y_{ij} $(i,\ j=1,\ 2,\ \ldots,\ s)$.

Deoarece, din (XIII.4,2) rezultă că funcția caracteristică a variabilelor $\bar{x}_j$ $(j=1,\ 2,\ \ldots,\ s)$, y_{ij} $(i,\ j=1,\ 2,\ \ldots,\ s)$ este egală cu funcția caracteristică a variabilelor x_j $(j=1,\ 2,\ \ldots,\ s)$ înmulțită cu funcția caracteristică a variabilelor y_{ij} $(i,\ j=1,\ 2,\ \ldots,\ s)$, tragem

concluzia că variabilele $\bar{x}_j$ $(j=1, 2, \ldots, s)$ şi y_{ij} $(i, j=1, 2, \ldots, s)$ sînt independente între ele. Se păstrează, prin urmare, şi pentru spaţiul cu s dimensiuni proprietatea pe care am demonstrat-o pe dreaptă: mediile de selecţie sînt independente de momentele centrate de selecţie de ordinul al doilea. Putem merge însă şi mai departe şi să determinăm în mod exact distribuţiile variabilelor $\bar{x}_j$ $(j=1, 2, \ldots, s)$ şi y_{ij} $(i, j=1, 2, \ldots, s)$.

Rezultă imediat, ţinînd seama de formulele stabilite la distribuţia normală cu mai multe dimensiuni, că

$$\varphi_1(t) = e^{-\frac{1}{2n} t'At}.$$

Variabilele $\bar{x}_j$ $(j=1, 2, \ldots, s)$ urmează, prin urmare, o distribuţie normală, avînd aceeaşi medie ca şi distribuţia teoretică şi momentele centrate de ordinul al doilea egale respectiv cu momentele centrate de ordinul al doilea ale distribuţiei teoretice, înmulţite cu factorul $\frac{1}{2}$.

Regăsim astfel teorema de la punctul 1.

Pe de altă parte

$$\varphi_2(nt) =$$

$$= \frac{|A|^{\frac{n-1}{2}}}{(2\pi)^{\frac{(n-1)s}{2}}} \iint \ldots \int e^{i \sum_{h,k} \delta_{hk} t_{hk} \sum_{j=2}^{n} v_{hj} v_{kj} - \frac{1}{2} \sum_{i,j} a_{ij} \sum_{h=2}^{n} v_{ih} v_{jh}} \prod dv_{ij}$$

$$(i=1, 2, \ldots, s) \quad (j=2, 3, \ldots, n)$$

reprezintă, prin definiţie, funcţia caracteristică a distribuţiei Wishart cu $n-1$ grade de libertate, avînd funcţia de distribuţie

$$\frac{|A|^{\frac{n-1}{2}}}{A_{n-1}} |X|^{\frac{n-s-2}{2}} \cdot e^{-\frac{1}{2} \sum_{i,j} a_{ij} x_{ij}}.$$

Rezultă printr-un calcul simplu că $\varphi_2(t)$ reprezintă funcția caracteristică a distribuției cu funcția de frecvență

$$\frac{|A|^{\frac{n-1}{2}}}{A_{\frac{n-1}{2}}} n^{\frac{(n-1)s}{2}} |X|^{\frac{n-s-2}{2}} \cdot e^{-\frac{1}{2}\sum_{i,j=1}^{s} a_{ij}x_{ij}}.$$

Am demonstrat astfel următoarea.

Teoremă. *Fie*

$$x_1, x_2, \ldots, x_n$$

o selecție de n variabile independente, fiecare fiind distribuită după legea normală s-dimensională de rangul s

$$\frac{|A|^{\frac{1}{2}}}{(2\pi)^{\frac{s}{2}}} \cdot e^{-\frac{1}{2}x'Ax}.$$

Dacă notăm

$$x_k = (x_{1k}, x_{2k}, \ldots, x_{sk}) \qquad (k=1, 2, \ldots, n),$$

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij} \qquad (j=1, 2, \ldots, s),$$

$$y_{ij} = \frac{1}{n}\sum_{h=1}^{n} (x_{ih}-\bar{x}_i)\ (x_{jh}-\bar{x}_j) \qquad (i, j=1, 2, \ldots, s),$$

avem următoarele proprietăți:

1) *grupurile de variabile* $\bar{x}_j$ $(j=1, 2, \ldots, s)$ *și* y_{ij} $(i, j=1, 2, \ldots, s)$ *sînt independente între ele;*

2) *variabilele* $\bar{x}_j$ $(j=1, 2, \ldots, s)$ *urmează o lege normală cu s dimensiuni, avînd aceeași medie ca distribuția teoretică și momentele centrate de ordinul al doilea respectiv egale cu momentele centrate de ordinul al doilea ale distribuției teoretice înmulțite cu* $\frac{1}{n}$;

3) *momentele centrate de selecție de ordinul al doilea, notate prin* y_{ij} $(i, j=1, 2, \ldots, s)$ *urmează o lege Wishart cu* $n-1$ *grade de libertate cu funcția de frecvență*

$$\frac{|A|^{\frac{n-1}{2}}}{A_{\frac{n-1}{2}}} n^{\frac{(n-1)s}{2}} |X|^{\frac{n-s-2}{2}} \cdot e^{-\frac{1}{2}\sum a_{ij}x_{ij}}.$$

Această teoremă este fundamentală pentru studiul selecţiilor dintr-o populaţie normală cu s dimensiuni.

Ea ne dă distribuţia mixtă a momentelor de selecţie de primele două ordine şi de aceea ne permite determinarea distribuţiilor funcţiilor de aceste momente.

Avem astfel la îndemînă o metodă pentru a determina distribuţiile de selecţie ale tuturor funcţiilor de care ne-am ocupat în teoria corelaţiei.

3. Am văzut la § 2 din acest capitol că raportul dintre valoarea medie a abaterii selecţiei şi momentul centrat de selecţie de ordinul al doilea se distribuie după o lege Student cu $n-1$ grade de libertate. Această teoremă se poate generaliza pentru o distribuţie teoretică cu s dimensiuni.

Am notat momentul centrat de selecţie de ordinul al doilea prin

$$y_{ij}=\frac{1}{n}\sum_{h=1}^{n}(x_{ih}-\bar{x}_i)\,(x_{jh}-\bar{x}_j) \qquad (i,\ j=1,\ 2,\ldots,\ s).$$

Să notăm prin z_{ij} momentul de selecţie de ordinul al doilea

$$z_{ij}=\frac{1}{n}\sum_{h=1}^{n}x_{ih}x_{jh} \qquad (i,\ j=1,\ 2,\ldots,\ s).$$

Avem

$$z_{ij}=y_{ij}+\bar{x}_i\bar{x}_j \qquad (i,\ j=1,\ 2,\ldots,\ s).$$

Folosind aceleaşi notaţii ca şi pînă acum, rezultă din teoria determinanţilor că

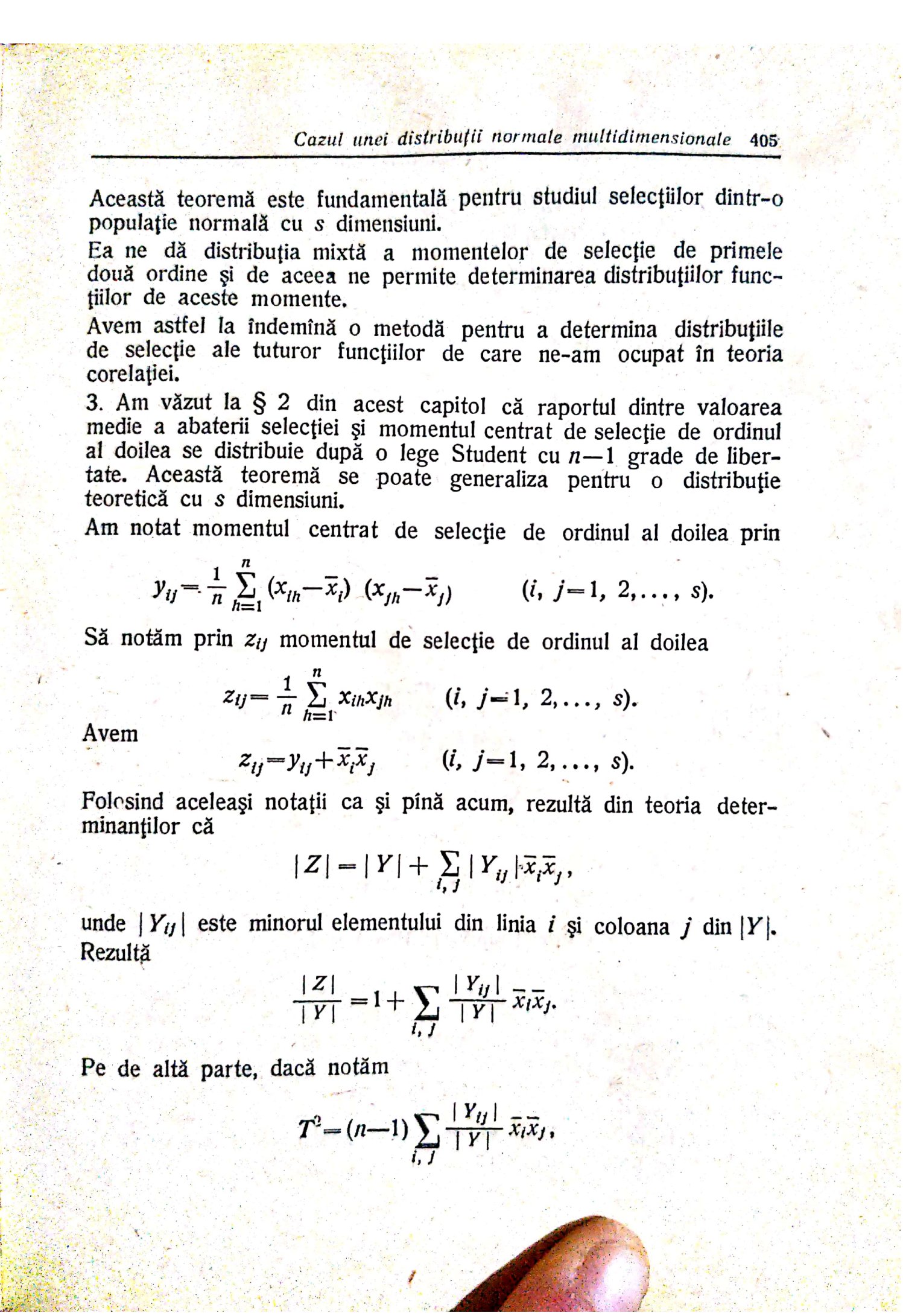
$$|Z|=|Y|+\sum_{i,\,j}|Y_{ij}|\bar{x}_i\bar{x}_j,$$

unde $|Y_{ij}|$ este minorul elementului din linia i şi coloana j din $|Y|$. Rezultă

$$\frac{|Z|}{|Y|}=1+\sum_{i,\,j}\frac{|Y_{ij}|}{|Y|}\bar{x}_i\bar{x}_j.$$

Pe de altă parte, dacă notăm

$$T^2=(n-1)\sum_{i,\,j}\frac{|Y_{ij}|}{|Y|}\bar{x}_i\bar{x}_j,$$



deoarece variabilele y_{ij} $(i, j=1, 2, \ldots, s)$ urmează o lege Wishart cu $n-1$ grade de libertate, iar variabilele $\bar{x}_i$ $(i=1, 2, \ldots, s)$, fiind independente de variabilele y_{ij}, urmează o distribuţie normală cu media zero şi cu aceeaşi matrice fundamentală A, care intervine şi în distribuţia Wishart a variabilelor y_{ij}, reiese că T urmează legea Hotelling cu $n-1$ grade de libertate, avînd funcţia de frecvenţă

$$\frac{2\,\Gamma\left(\frac{n}{2}\right)}{(n-1)^{\frac{s}{2}}\,\Gamma\left(\frac{n-s}{2}\right)\Gamma\left(\frac{s}{2}\right)} \cdot \frac{x^{k-1}}{\left(1+\frac{x^2}{n-1}\right)^{\frac{s}{2}}} \qquad (s>0).$$

Din cele de mai sus, se vede că putem determina pe T şi din relaţia

$$\frac{|Z|}{|Y|} = \frac{1}{1+\frac{T^2}{n-1}},$$

unde $|Z|$ este determinantul format cu momentele de selecţie de ordinul al doilea, iar $|Y|$ determinantul format cu momentele centrate de selecţie de ordinul al doilea.

Capitolul XIV

Sondaje cu o fază

§ 1. Consideraţii generale

Metoda sondajelor sau a eşantioanelor este o metodă foarte des aplicată în practica statisticii şi constituie, de fapt, un caz particular al teoriei generale a selecţiei de care ne-am ocupat în capitolul precedent. Prin această metodă se urmăreşte, de obicei, estimarea mediei unei colectivităţi statistice cu ajutorul unei selecţii. Cazul cel mai simplu este atunci cînd media a a distribuţiei teoretice nu este cunoscută şi se cunoaşte funcţia ei de frecvenţă, în care a intră ca un parametru.

Scanned with OKEN Scanner

Se poate întîmpla, însă, ca distribuția teoretică să nu fie unică și să existe mai multe distribuții teoretice dacă, de exemplu, în fiecare dintre cele k regiuni

$$D_1, \quad D_2, \ldots, D_k$$

din spațiul R_s este o altă distribuție.

Atunci, se pune problema, cum trebuie să facem selecția din toate aceste regiuni pentru a estima media care ne interesează? In cazul cînd distribuția dată este finită, metoda selecției se poate aplica în două moduri.

Primul mod, corespunzător schemei probabilistice a lui Bernoulli, constă în faptul că elementele care au apărut în selecție pot apare de mai multe ori.

In acest caz, variabilele selecției sînt independente între ele și dau naștere unei distribuții după modelul urnei lui Bernoulli.

Al doilea mod corespunde distribuției hipergeometrice, concretizat printr-o urnă din care facem extracții succesive, fără să punem bila extrasă înapoi în urnă. In acest caz, variabilele selecției sînt dependente între ele.

O selecție finită făcută cu scopul arătat mai sus, poartă numele de *sondaj* sau *eșantion*.

Unele condiții de ordin practic permit să se înțeleagă mai bine deosebirile dintre diferitele feluri de sondaje.

Să presupunem, astfel, că prin metoda sondajelor sau a eșantioanelor ne interesează determinarea procentului de analfabeți, pe baza datelor unui recensămînt general dintr-o țară, din care prelucrăm numai un material de 5%, luat la întîmplare din volumul total. Luînd toți locuitorii care fac parte din acest sondaj numai din mediul urban, rezultatul obținut ar fi, firește, eronat, dacă în realitate există o diferență mare între procentele de analfabeți din mediul urban și cel rural. La fel am ajunge la un rezultat eronat, dacă tot sondajul ar fi efectuat numai din populația rurală. Procentul de 5% din populația totală, care ne dă mărimea sondajului, va trebui luat și din populația urbană și din cea rurală. Vom lua, deci, $\alpha\%$ la întîmplare, locuitori din mediul urban și $\beta\%$ la întîmplare, locuitori din mediul rural, astfel ca $\alpha+\beta=5$. Vom face, prin urmare, un sondaj stratificat cu două straturi și anume, un strat în mediul urban și un strat în mediul rural. Coeficienții α și β trebuie să-i alegem în așa fel, încît rezultatul obținut să fie cît mai aproape de realitate.

Vom spune că un sondaj A de mărime n, $A(n)$, în baza căruia estimăm parametrul p prin funcţia de selecţie $\bar{p}_A$ este mai *eficient* decît sondajul $B(n)$, tot de mărime n, în baza căruia estimăm parametrul p prin funcţia de selecţie $\bar{p}_B$, dacă avem următoarele relaţii

$$M(\bar{p}_A)=p, \quad M(\bar{p}_B)=p, \quad D(\bar{p}_A)<D(\bar{p}_B).$$

Variabilele

$$x_1, \quad x_2, \ldots, x_n$$

ale sondajului poartă numele de *unităţi* ale sondajului sau eşantionului. De unde rezultă că un sondaj de mărime n este compus din n unităţi.

§ 2. Sondaje pentru o distribuţie infinită

1. Să considerăm o distribuţie unidimensională a cărei funcţie de repartiţie este $F(x)$. Atunci, valoarea medie a acestei distribuţii este

$$m=\int_{-\infty}^{+\infty} x\,dF(x),$$

care presupunem că există.

Dacă facem n selecţii independente

$$x_1, \quad x_2, \ldots, x_n,$$

atunci media de selecţie

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n} x_i$$

este, după cum se ştie, o estimaţie absolut corectă a mediei teoretice m.

2. Să presupunem că distribuţia considerată nu este unică şi, ca atare, există k straturi

$$D_1, \quad D_2, \ldots, D_k,$$

unde, în fiecare dintre aceste straturi, este o altă distribuţie a căror funcţie de repartiţie o notăm respectiv cu

$$F_1(x), \quad F_2(x), \ldots, F_k(x).$$

Scanned with OKEN Scanner

De asemenea, notăm respectiv prin

$$p_1, \quad p_2, \ldots, p_k \qquad \left(\sum_{i=1}^{k} p_i = 1\right)$$

probabilitatea ca vectorul x să cadă într-unul din straturile D_i, prin

$$m_i = \int_{-\infty}^{+\infty} x dF_i(x)$$

şi prin

$$\sigma_i^2 = \int_{-\infty}^{+\infty} (x - m_i)^2 dF(x) \qquad (i = 1, \quad 2, \ldots, k)$$

respectiv, media şi dispersia distribuţiei teoretice corespunzătoare stratului D_i, care medii şi dispersii admitem că există.

Ne propunem să estimăm media

$$m = \sum_{i=1}^{k} p_i m_i$$

pe baza unui sondaj de mărime n, făcut la întîmplare din toate straturile regiunii

$$D = D_1 \cup D_2 \cup \ldots \cup D_k,$$

straturile D_i fiind disjuncte două cîte două, adică

$$D_i \cap D_j = \Phi \qquad (i \neq j)$$

unde Φ înseamnă mulţimea vidă.

Atunci, probabilitatea ca $x \in D_i$, adică probabilitatea ca vectorul x luat la întîmplare să cadă în stratul D_i şi să avem în acelaşi timp $x < y$ este

$$P(x \in D_i \cap x < y) = P(x \in D_i) \cdot P(x < y \mid x \in D_i) = p_i F_i(y),$$

ţinînd seama de notaţiile noastre.

De aici rezultă că funcţia de repartiţie a variabilei vectoriale x luată la întîmplare din toată regiunea D este, după teorema probabilităţilor totale,

$$F(y) = P(x < y) = \sum_{i=1}^{k} p_i F_i(y). \qquad \text{(XIV.2,1)}$$

Efectuînd un sondaj de mărime n

$$x_1, \quad x_2, \ldots, x_n$$

avem pentru orice $x_j\,(j=1, 2, \ldots, n)$

$$M(x_j)=\int_{-\infty}^{+\infty} x dF(x)=\int_{-\infty}^{+\infty} x d\left(\sum_{i=1}^{k} p_i F_i(x)\right)=\sum_{i=1}^{k} p_i \int_{-\infty}^{+\infty} x dF_i(x)=$$

$$=\sum_{i=1}^{k} p_i m_i=m,$$

avînd în vedere relaţia (XIV.2,1).
Pe de altă parte,

$$M(\bar{x})=\frac{1}{n} M\left(\sum_{j=1}^{n} x_j\right)=\frac{1}{n}\sum_{j=1}^{n} M(x_j)=m, \qquad \text{(XIV.2,2)}$$

iar

$$D^2(\bar{x})=M(\bar{x}-m)^2=M(\bar{x})^2-m^2=\frac{1}{n^2} M\left(\sum_{j=1}^{n} x_j\right)^2-m^2=$$

$$=\frac{1}{n^2} M\left(\sum_{j=1}^{n} x_j^2+2\sum_{j\neq k} x_j x_k\right)-m^2=$$

$$=\frac{1}{n^2}\sum_{j=1}^{n} M(x_j^2)+\frac{2}{n^2}\sum_{j\neq k} M(x_j) M(x_k)-m^2=$$

$$=\frac{1}{n^2}\sum_{j=1}^{n}\int_{-\infty}^{+\infty} x_i^2 dF(x)+\frac{n-1}{n} m^2-m^2=\frac{1}{n}\sum_{i=1}^{k} p_i \int_{-\infty}^{+\infty} x^2 dF_i(x)-\frac{m^2}{n}=$$

$$=\frac{1}{n}\sum_{i=1}^{n} p_i(\sigma_i^2+m_i^2)-\frac{m^2}{n}=\frac{\sum_{i=1}^{n} p_i\sigma_i^2+\sum_{i=1}^{k} p_i(m_i-m)^2}{n}. \qquad \text{(XIV.2,3)}$$

Din expresia lui $D^2(\bar{x})$ deducem

$$\lim_{n\to\infty} D^2(\bar{x})=\lim_{n\to\infty}\frac{\sum_{i=1}^{k} p_i\sigma_i^2+\sum_{i=1}^{k} p_i(m_i-m)^2}{n}=0. \qquad \text{(XIV.2.4)}$$

Prin urmare, relațiile (XIV.2,2) și (XIV.2,4) ne arată că, dacă sondajul de mărime n

$$x_1, \quad x_2, \ldots, x_n$$

este luat la întîmplare din toate straturile regiunii

$$D=\bigcup_{i=1}^{k} D_i \qquad \left(D_i \bigcap_{i\neq j} D_j=\Phi\right),$$

atunci, media aritmetică

$$\bar{x}=\frac{1}{n}\sum_{j=1}^{n} x_j$$

estimează media generală

$$m=\sum_{i=1}^{k} p_i m_i.$$

Această estimație, după cum se observă, este absolut corectă.

3. Să presupunem, acum, că sondajul de mărime n este compus din k sondaje de mărimi respectiv egale cu

$$n_1, \quad n_2, \ldots, n_k \qquad \left(\sum_{i=1}^{k} n_i=n\right)$$

efectuate din straturile corespunzătoare

$$D_1, \quad D_2, \ldots, D_k.$$

In baza acestui sondaj de mărime n

$$n_1+n_2+\ldots+n_k=n,$$

ne propunem să găsim o funcție de estimație pentru medie

$$m=\sum_{i=1}^{k} p_i m_i.$$

Atunci, efectuînd sondajul de mărime n, fie respectiv

$$\begin{aligned} &x_{11}, \quad x_{21}, \ldots, x_{n_1 1}, \\ &x_{12}, \quad x_{22}, \ldots, x_{n_2 2}, \\ &\cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \\ &x_{1k}, \quad x_{2k}, \ldots, x_{n_k k}, \end{aligned}$$

unităţile corespunzătoare straturilor

$$D_1, \quad D_2, \ldots, D_k.$$

Considerînd media aritmetică

$$\bar{x} = \frac{1}{n}\left(\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}\right), \qquad \text{(XIV.2,5)}$$

avem

$$M(\bar{x}) = \frac{1}{n} M\left(\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}\right) = \sum_{j=1}^{k} \frac{\bar{n}_j}{n} m_j,$$

dacă introducem notaţia

$$\bar{x}_j = \frac{1}{n_j}\left(\sum_{i=1}^{n_j} x_{ij}\right)$$

şi ţinem seama că $M(\bar{x}_j) = n_j$, pentru orice $j = 1, 2, \ldots, k$.
Pentru a avea

$$M(\bar{x}) = m$$

sau

$$\sum_{j=1}^{k} \frac{n_j}{n} m_j = \sum_{j=1}^{k} p_i m_i,$$

trebuie să aibă loc egalitatea

$$\frac{n_j}{n} = p_i,$$

adică

$$n_j = [np_j] \qquad (j = 1, 2, \ldots, k).$$

Cu alte cuvinte, numărul unităţilor dintr-un strat oarecare D_j trebuie să fie egal cu cel mai mare întreg din np_j, adică $[np_j]$.
Un asemenea sondaj se numeşte *sondaj reprezentativ sau proporţional*, iar funcţia de selecţie (XIV.2,5), corespunzătoare acestui sondaj, o notăm cu $\bar{x}_r$.

Scanned with OKEN Scanner

Dispersia distribuţiei variabilei x_r este

$$D^2(\bar{x}_r)=M\left(\bar{x}-\sum_{j=1}^{k}\frac{n_j}{n}m_j\right)^2=M\left(\sum_{j=1}^{k}\frac{n_j}{n}\bar{x}_j-\sum_{j=1}^{k}\frac{n_j}{n}m_j\right)^2=$$

$$=M\left[\sum_{j=1}^{k}\frac{n_j}{n}(\bar{x}_j-m_j)\right]^2,$$

folosind notaţia introdusă.

Sau, observînd că variabilele $\bar{x}_j-m_j$ şi $\bar{x}_l-m_l$ sînt independente între ele, obţinem

$$D^2(\bar{x}_r)=\sum_{j=1}^{k}\frac{n_j^2}{n^2}M(\bar{x}_j-m_j)^2=\sum_{j=1}^{k}\frac{n_j^2}{n^2}\cdot\frac{\sigma_j^2}{n_j}=\sum_{j=1}^{k}\frac{1}{n}\frac{n_j}{n}\sigma_j^2=\frac{1}{n}\sum_{j=1}^{k}p_j\sigma_j^2, \quad \text{(XIV.2,6)}$$

ţinînd seama de expresia dispersiei mediei de selecţie $M(\bar{x}_j-m_j)^2=\frac{\sigma_j^2}{n_j}$.

Teoremă. *Sondajul reprezentativ este mai eficient decît sondajul nestratificat.*

Demonstraţia teoremei revine la a arăta că

$$D^2(\bar{x}_r)\leqslant D^2(\bar{x}),$$

unde am însemnat prin $D^2(\bar{x}_r)$ şi $D^2(x)$, dispersiile corespunzătoare celor două sondaje.

Inlocuind aceste dispersii prin expresiile lor din (XIV.2,6) şi (XIV.2,3), inegalitatea de mai sus devine

$$\sum_{j=1}^{k}p_j\sigma_j^2\leqslant\sum_{j=1}^{k}p_j\sigma_j^2+\sum_{j=1}^{k}p_j(m_j-m)^2,$$

ceea ce este evident.

Egalitatea are loc numai dacă $m_j=m$ $(p_j\neq 0)$ şi atunci sondajul reprezentativ se reduce la un sondaj nestratificat.

In cazul sondajului reprezentativ am presupus că probabilităţile $p_j\,(j=1,\ 2,\ \ldots,k)$ sînt cunoscute. Dacă, în plus, presupunem cu-

Scanned with OKEN Scanner

noscute şi dispersiile $\sigma_j^2\ (j=1,\ 2,\ \ldots,k)$ nu putem să ajungem la un sondaj mai eficient.

Considerînd acum media

$$\bar{x}=\sum_{j=1}^{k} p_j\bar{x}_j\,, \tag{XIV.2,7}$$

putem scrie

$$M(\bar{x})=\sum_{j=1}^{k} p_jM(\bar{x}_j)=\sum_{j=1}^{k} p_jm_j=m,$$

oricare ar fi unităţile sondajului din fiecare strat al regiunii D.

Avem

$$D^2(\bar{x})=M\left(\sum_{j=1}^{k} p_j\bar{x}_j-m\right)^2=M\left(\sum_{j=1}^{k} p_j\bar{x}_j-\sum_{j=1}^{k} p_jm_j\right)^2=$$

$$=M\left[\sum_{j=1}^{k} p_j(\bar{x}_j-m_j)\right]^2=\sum_{j=1}^{k} p_j^2M(\bar{x}_j-m_j)^2=\sum_{j=1}^{k} p_j^2D^2(\bar{x}_j)=\sum_{j=1}^{k}\frac{p_j^2\sigma_j^2}{n_j}, \tag{XIV.2,8}$$

deoarece variabilele $\bar{x}_j-m_j$ şi $\bar{x}_l-m_l$ sînt independente între ele, iar $D^2(\bar{x}_j)=\dfrac{\sigma_j^2}{n_j}$.

Atunci, dacă n_j este cel mai mic număr dintre toate numerele

$$n_1,\ n_2,\ \ldots,\ n_k$$

şi $n_j\to\infty$, se vede că

$$\lim_{n_j\to\infty} D^2(\bar{x})=\lim_{n_j\to\infty}\sum_{j=1}^{k}\frac{p_j^2\sigma_j^2}{n_j}=0$$

şi, prin urmare, funcţia $\bar{x}$ definită de expresia (XIV.2,7) este o estimaţie absolut corectă a mediei teoretice m.

Mai departe, ne propunem să determinăm numerele $n_j\ (j=1,2,\ldots,k)$ astfel ca eficienţa sondajului să fie maximă, iar suma

$$n_1+n_2+\ldots+n_k$$

să fie egală cu n.

Aceasta înseamnă, potrivit definiţiei sondajului mai eficient, că trebuie să determinăm numerele $n_j\,(j=1,\ 2,\ldots,k)$ în aşa fel, încît expresia

$$D^2(\bar{x})=\sum_{j=1}^{k}\frac{p_j^2\sigma_j^2}{n_j}$$

să fie minimă, îndeplinind condiţia

$$n_1+n_2+\ldots+n_k=n.$$

Pentru aflarea acestui minim, vom aplica metoda multiplicatorilor lui Lagrange. Aşadar, scriem

$$F(n_1,\ n_2,\ldots,n_k)=\sum_{j=1}^{k}\frac{p_j^2\sigma_j^2}{n_j}+\lambda\left(\sum_{j=1}^{k}n_j-n\right).$$

Anulînd derivatele parţiale de ordinul întîi în raport cu n_j, obţinem

$$\frac{\partial F(n_1,\ n_2,\ldots,n_k)}{\partial n_j}=\frac{p_j^2\sigma_j^2}{n_j^2}+\lambda=0\qquad(j=1,\ 2,\ldots,k)$$

sau

$$\frac{p_j\sigma_j}{n_j}=\sqrt{\lambda},\qquad(j=1,\ 2,\ldots,k)$$

luînd radicalul cu semnul plus, deoarece p_j, σ_j şi n_j sînt toate pozitive.

Din proprietatea elementară a unui şir de rapoarte egale, deducem

$$\sqrt{\lambda}=\frac{\sum_{j=1}^{k}p_j\sigma_j}{n},$$

de unde rezultă că

$$\frac{p_j\sigma_j}{n_j}=\frac{\sum_{j=1}^{k}p_j\sigma_j}{n},$$

adică

$$n_j=\frac{np_j\sigma_j}{\sum_{j=1}^{k}p_j\sigma_j}\qquad(j=1,\ 2,\ldots,k).\qquad\text{(XIV.2,9)}$$

Am determinat, astfel, numerele n_j pentru care eficienţa sondajului este maximă.

Un asemenea sondaj poartă numele de *sondaj optim*, iar funcţia de selecţie (XIV.2,7) care răspunde la această problemă de minim, adică corespunzătoare sondajului optim, o însemnăm prin $\bar{x}_0$.

Dispersia distribuţiei lui $\bar{x}_0$ este

$$D^2(\bar{x}_0)=\sum_{j=1}^{k} p_j^2\sigma_j^2 \frac{\sum_{j=1}^{k} p_j\sigma_j}{np_j\sigma_j}=\frac{1}{n}\left(\sum_{j=1}^{k} p_j\sigma_j\right)^2. \quad \text{(XIV.2,10)}$$

Teoremă. *Sondajul optim este mai eficient decît sondajul reprezentativ.*

Demonstraţia teoremei nu rezultă din proprietatea de minim de mai sus, deoarece, după cum am văzut, sondajul reprezentativ s-a calculat în ipoteza cunoaşterii probabilităţilor p_j, iar sondajul optim în ipoteza cunoaşterii probabilităţilor p_j şi a dispersiilor σ_j^2. Deci sondajul reprezentativ nu este un caz particular al sondajului optim.

Demonstraţia teoremei rezultă, totuşi, imediat.

Pentru aceasta, trebuie să arătăm că

$$D^2(\bar{x}_0)\leqslant D^2(\bar{x}_r)$$

sau, înlocuind dispersiile prin expresiile lor din (XIV.2,10) şi (XIV.2,6)

$$\left(\sum_{j=1}^{k} p_j\sigma_j\right)^2 \leqslant \sum_{j=1}^{k} p_j\sigma_j^2 \sum_{j=1}^{k} p_j \qquad \left(\sum_{j=1}^{k} p_j=1\right).$$

Această inegalitate este, însă, un caz particular al inegalităţii lui Schwarz. In adevăr notînd

$$A_j=\sqrt{p_j}, \qquad B_j=\sqrt{p_j}\cdot\sigma_j,$$

obţinem inegalitatea lui Schwarz

$$\left(\sum_{j=1}^{k} A_jB_j\right)^2 \leqslant \sum_{j=1}^{k} A_j^2 \sum_{j=1}^{k} B_j^2.$$

Teorema este, astfel, demonstrată.

Scanned with OKEN Scanner

In practica metodei sondajelor se mai pun şi alte probleme, ca de pildă, acelea care se leagă de latura lor economică. Astfel, se poate fixa totalul cheltuielilor sondajului, ca să fie egal cu o sumă oarecare C. In ceea ce priveşte cheltuielile de la strat la strat, ele, în general, variază. Aşa, de exemplu, un sondaj în mediul rural ar putea fi mai scump decît un sondaj în mediul urban. De aceea, notăm cu c_j cheltuiala corespunzătoare unei unităţi din stratul D_j. Atunci, avem evident

$$C=\sum_{j=1}^{k} n_j c_j.$$

Presupunînd că se cunosc probabilităţile p_j, dispersiile σ_j^2, totalul cheltuielilor C şi cheltuielile c_j pe o singură unitate, ne putem pune problema determinării numărului unităţilor $n_j\,(j=1,\ 2,\ \ldots,\ k)$ din fiecare strat, astfel ca dispersia funcţiei de selecţie

$$\bar{x}=\sum_{j=1}^{k} p_j \bar{x}_j$$

să fie minimă, condiţionată în cazul de faţă de egalitatea

$$C=n_1 c_1+n_2 c_2+\ldots+n_k c_k.$$

Dispersia acestei funcţii de selecţie am calculat-o mai înainte şi este dată de expresia (XIV.2,8), adică

$$D^2(\bar{x})=\sum_{j=1}^{k}\frac{p_j^2\sigma_j^2}{n_j}.$$

După metoda multiplicatorilor lui Lagrange, punem

$$F(n_1,\ n_2,\ \ldots,n_k)=\sum_{j=1}^{k}\frac{p_j^2\sigma_j^2}{n_j}+\lambda\left(\sum_{j=1}^{k} n_j c_j-C\right).$$

Pentru a obţine un minim, trebuie ca

$$\frac{\partial F(n_1,\ n_2,\ \ldots,n_k)}{\partial n_j}=0 \qquad (j=1,\ 2,\ \ldots,\ k)$$

sau, ceea ce este acelaşi lucru,

$$\frac{p_j^2\sigma_j^2}{n_j^2}=\lambda c_j,$$

de unde deducem

$$n_j = \frac{p_j \sigma_j}{\sqrt{\lambda c_j}} \qquad (j=1,\ 2,\ \ldots,\ k), \qquad \text{(XIV.2,11)}$$

luînd ca şi în problema de minim precedentă, radicalul cu semnul plus, deoarece $n_j > 0$.
Inmulţind această egalitate cu c_j şi însumînd după indicele j, de la 1 la k, obţinem

$$C = \frac{\sum_{j=1}^{k} p_j \sigma_j \sqrt{c_j}}{\sqrt{\lambda}},$$

adică

$$\sqrt{\lambda} = \frac{\sum_{j=1}^{k} p_j \sigma_j \sqrt{c_j}}{C}.$$

Acum, înlocuind pe $\sqrt{\lambda}$ în (XIV.2,11), găsim

$$n_j = \frac{C p_j \sigma_j}{\sqrt{c_j} \sum_{j=1}^{k} p_j \sigma_j \sqrt{c_j}}. \qquad \text{(XIV.2,12)}$$

Această expresie, care este valabilă oricare ar fi $j=1,\ 2,\ \ldots,\ k$, ne determină numărul unităţilor din fiecare strat în condiţiile puse de problema noastră.
In particular, cînd avem

$$c_1 = c_2 = \ldots = c_k = c,$$

rezultă

$$C = c \sum_{j=1}^{k} n_j = nc$$

şi, deci, relaţia (XIV.2,12) devine

$$n_j = \frac{n p_j \sigma_j}{\sum_{j=1}^{k} p_j \sigma_j}.$$

Regăsim, astfel, expresia (XIV.2,9) a numărului unităţilor sondajului optim.

Scanned with OKEN Scanner

In sfîrşit, însemnînd cu $\bar{x}_c$ funcţia de selecţie corespunzătoare, avem

$$D^2(\bar{x}_c)=\sum_{j=1}^{k}\frac{p_j^2\sigma_j^2}{n_j}=\sum_{j=1}^{k}p_j^2\sigma_j^2\frac{\sqrt{c_j}\sum_{j=1}^{k}p_j\sigma_j\sqrt{c_j}}{Cp_j\sigma_j}=\frac{1}{C}\left(\sum_{j=1}^{k}p_j\sigma_j\sqrt{c_j}\right)^2.$$

Această expresie ne dă eficienţa sondajului.
Se mai poate pune şi o altă problemă de natură economică. De exemplu, se poate fixa de la început eficienţa sondajului şi se cere să se determine numărul unităţilor n_j $(j=1,2,\ldots,k)$ din fiecare strat D_j al regiunii D în aşa fel, încît totalul cheltuielilor C să fie minim.
Prin urmare, ni se dă relaţia

$$\sum_{j=1}^{k}\frac{p_j^2\sigma_j^2}{n_j}=a, \qquad \text{(XIV.2,13)}$$

care exprimă faptul că dispersia funcţiei de estimaţie, în baza căreia apreciem eficienţa sondajului, este egală cu o constantă dată a.
In acest caz, va trebui să determinăm numărul unităţilor n_j din fiecare strat D_j, astfel ca expresia

$$\sum_{j=1}^{k}n_jc_j$$

să fie minimă, condiţionată de relaţia (XIV.2,13).
Procedînd ca şi în problema precedentă, după metoda multiplicatorilor lui Lagrange, vom scrie

$$F(n_1,n_2,\ldots,n_k)=\sum_{j=1}^{k}n_jc_j+\lambda\left(\sum_{j=1}^{k}\frac{p_j^2\sigma_j^2}{n_j}-a\right),$$

de unde obţinem, anulînd derivata parţială de ordinul întîi în raport cu n_j a funcţiei $F(n_1,n_2,\ldots,n_k)$

$$\frac{\partial F(n_1,n_2,\ldots,n_k)}{\partial n_j}=c_j-\lambda\frac{p_j^2\sigma_j^2}{n_j^2}=0 \qquad (j=1,2,\ldots,k),$$

adică

$$n_j=\frac{p_j\sigma_j\sqrt{\lambda}}{\sqrt{c_j}}, \qquad \text{(XIV.2,14)}$$

radicalul fiind luat cu semnul plus, deoarece $n_j>0$.

Scanned with OKEN Scanner

Inlocuind pe n_j, relaţia (XIV.2,13) devine

$$\frac{1}{\sqrt{\lambda}}\sum_{j=1}^{k}\frac{p_j^2\sigma_j^2\sqrt{c_j}}{p_j\sigma_j}=a$$

sau

$$\sqrt{\lambda}=\frac{\sum_{j=1}^{k}p_j\sigma_j\sqrt{c_j}}{a}.$$

Efectuind această substituţie în expresia (XIV.2,14), rezultă numărul căutat al unităţilor n_j

$$n_j=\frac{p_j\sigma_j\sum_{j=1}^{k}p_j\sigma_j\sqrt{c_j}}{a\sqrt{c_j}}.$$

Atunci, cheltuiala minimă este

$$C=\sum_{j=1}^{k}n_jc_j=\frac{1}{a}\left(\sum_{j=1}^{k}p_j\sigma_j\sqrt{c_j}\right)^2.$$

Se observă analogia între minimul acesta şi cel precedent, care diferă între ele prin factorul $\frac{1}{a}$, respectiv $\frac{1}{C}$.

§ 3. Sondaje dintr-o colectivitate finită

De cele mai multe ori, în practica sondajelor, colectivităţile întîlnite sînt finite şi putem proceda cu ele corespunzător schemei urnei lui Bernoulli. Fiecare unitate ipotetic extrasă rămîne de fapt în colectivitatea totală. In modul acesta, variabilele selecţiei pot fi independente între ele, iar fiecare dintre unităţile colectivităţii poate să apară de mai multe ori. Sîntem, deci, în cazul examinat în praragraful precedent, cînd am studiat sondajele dintr-o colectivitate infinită. De altfel, procedînd după schema urnei lui Bernoulli, succesiunea probelor de selecţie este infinită şi problema se reduce, astfel, la cea precedentă a sondajelor dintr-o populaţie infinită.

Examinăm aici, însă, problema sondajelor dintr-o colectivitate finită, cînd unităţile selecţionate dispar din colectivitatea generală. In acest caz, variabilele selecţiei sînt necesar dependente, iar o unitate nu poate să apară în şirul succesiv al selecţiilor decît o singură dată.

Scanned with OKEN Scanner

Aceasta ne conduce la o distribuţie hipergeometrică, a cărei schemă de probabilitate corespunde extracţiilor dintr-o urnă din care se fac extracţii succesive, fără să se pună bila extrasă înapoi în urnă, altminteri însă independente unele de altele.

Fie deci o colectivitate statistică A, reprezentată de N valori numerice

$$a_1, a_2, \ldots, a_N,$$

a cărei medie este

$$a = \frac{1}{N} \sum_{i=1}^{N} a_i.$$

Se pune problema estimării acestei medii cu ajutorul unui sondaj de mărime n

$$x_1, x_2, \ldots, x_n,$$

scoţind la întîmplare n elemente din colectivitatea A, fără să le restituim. Avem astfel un *sondaj simplu.*

Să imaginăm colectivitatea A împărţită în k straturi

$$A_1, A_2, \ldots, A_k,$$

fiecare dintre ele avînd respectiv elementele

$$a_{11}; a_{21}, \ldots, a_{N_1 1},$$

$$a_{12}, a_{22}, \ldots, a_{N_2 2},$$

$$\ldots \ldots \ldots \ldots$$

$$a_{1k}, a_{2k}, \ldots, a_{N_k k}.$$

Să efectuăm cîte

$$n_1, n_2, \ldots, n_k \qquad \left(\sum_{j=1}^{k} n_j = n\right)$$

extracţii din straturile respective $A_1, A_2, \ldots, A_{N_k}$, după acelaşi procedeu ca mai sus. Un sondaj făcut în aceste condiţii dă pentru straturile respective şirurile de valori

$$x_{11}, x_{21}, \ldots, x_{n_1 1},$$

$$x_{12}, x_{22}, \ldots, x_{n_2 2},$$

$$\ldots \ldots \ldots \ldots$$

$$x_{1k}, x_{2k}, \ldots, x_{n_k k}$$

şi poartă numele de *sondaj stratificat simplu.*

Dacă din anumite motive, nu facem sondaje din toate straturile $A_j\,(j=1,2,\ldots,k)$ ci ne limităm la $h<k$ straturi pe care, firește le alegem la întîmplare, sondajul se efectuează în două faze.

In prima fază, din totalitatea straturilor

$$A_1, A_2, \ldots, A_k$$

alegem la întîmplare succesiv un număr de $h<k$ straturi distincte. Colectivitatea din care extragem cele h straturi este formată din elementele $A_j\,(j=1,2,\ldots,k)$. Aceste elemente constituie *unităţile primare* ale sondajului. Fie $A'_1, A'_2, \ldots, A'_h$ cele h straturi în ordinea prezentării lor. Presupunem $A'_1=A_j$. In faza a doua, din totalitatea elementelor

$$a_{1j}, a_{2j}, \ldots, a_{N_jj}$$

ale stratului $A'_1=A_j\,(j=1,2,\ldots,k)$, efectuăm n_j extracţii la întîmplare, obţinînd selecţia

$$x_{1j}, x_{2j}, \ldots, x_{n_jj}.$$

Deoarece colectivitatea din care efectuăm selecţia este formată din elementele $a_{ij}\,(i=1,2,\ldots,N_h)$, $(j=1,2,\ldots,h)$, aceste elemente poartă numele de *unităţi secundare* ale sondajului.

Avem astfel un *sondaj simplu cu două faze*. In mod asemănător, se pot defini sondaje simple cu trei, patru şi mai multe faze.

In sondajul simplu cu trei faze, de exemplu, din rîndul unităţilor primare vom alege la întîmplare h dintre ele. Aceasta este prima fază a sondajului. Unităţile primare conţin fiecare un număr dat de unităţi secundare. Din unităţile primare selecţionate în prima fază, efectuăm un număr determinat de extracţii de unităţi secundare. *Aceasta este a doua fază a sondajului.* La rîndul lor, fiecare unitate secundară, conţine un număr determinat de unităţi terţiare. *Aceasta este a treia fază a sondajului.*

Este util să semnalăm şi un alt aspect al procedeului sondajelor. După cum reiese şi din exemplul următor, unităţile sondajului nu prezintă întotdeauna o singură caracteristică. Astfel, dacă de pildă, sondajul se referă la o colectivitate de o anumită categorie de profesionişti, putem considera drept caracteristici: gradul de pregătire în profesiune, vechimea în profesiune, salariul etc. Cerem în acest caz unui sondaj să ne conducă la o corelaţie între aceste caracteristici.

Scanned with OKEN Scanner

Rezultă de aici, cum, în practica sondajelor, prezintă interes considerarea mai multor caracterstici pentru fiecare unitate a colectivităţii, precum şi stabilirea relaţiilor dintre ele.

§ 4. Sondajul simplu

1. Să considerăm colectivitatea A, ale cărei unităţi prezintă fiecare o singură caracteristică, înţelegînd prin aceasta, că unităţii de rang j îi corespunde singura valoare a_j $(j=1, 2, \ldots, N)$ şi să efectuăm un sondaj simplu de mărime n

$$x_1, x_2, \ldots, x_n.$$

Introducem notaţiile

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j,$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(a_i - a)^2, \qquad \text{(XIV.4,1)}$$

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(a_i - a)^2.$$

In ipoteza că fiecărui element a_j îi este atribuită frecvenţa $\frac{1}{N}$, se vede că σ^2 reprezintă dispersia distribuţiei A.

Teoremă. *Sondajele simple de unităţi care prezintă o singură caracteristică ne dau relaţiile*

$$M(\bar{x}) = a,$$

$$D^2(\bar{x}) = (1-f)\frac{s^2}{n},$$

unde $f = \frac{n}{N}$ $(n < N)$. *Deci media de selecţie*

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j$$

este o estimaţie absolut corectă a mediei teoretice

$$a = \frac{1}{N}\sum_{i=1}^{N} a_i.$$

Demonstrația teoremei este imediată. In adevăr, din studiul distribuției hipergeometrice (cap. IX, § 6) se știe că

$$M(\bar{x})=a,$$

$$D^2(\bar{x})=\frac{N-n}{N-1}\cdot\frac{\sigma^2}{n}. \qquad \text{(XIV.4,2)}$$

Eliminînd pe $\sum_{i=1}^{N}(a_i-a)^2$ între ultimele două relații din (XIV.4,1), obținem

$$\sigma^2=\frac{N-1}{N}s^2.$$

Inlocuind pe σ^2, relațiile (XIV.4,2) devin

$$M(\bar{x})=a,$$

$$D^2(\bar{x})=(1-f)\frac{s^2}{n}.$$

Se vede că, pentru n suficient de mare, $D^2(\bar{x})\to 0$. Teorema este demonstrată.

Teoremă. *Sondajele simple ale căror unități prezintă o singură caracteristică dau egalitatea*

$$M\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2\right]=s^2.$$

Pentru demonstrație, plecăm de la identitatea

$$\sum_{i=1}^{n}(x_i-\bar{x})^2=\frac{n-1}{n}\sum_{i=1}^{n}x_i^2-\frac{2}{n}\sum_{\substack{i,j=1\\ i\neq j}}^{n}x_ix_j. \qquad \text{(XIV.4,3)}$$

De unde rezultă

$$M\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2\right]=\frac{1}{n}\sum_{i=1}^{n}M(x_i^2)-\frac{2}{n(n-1)}\sum_{\substack{i,j=1\\ i\neq j}}^{n}M(x_ix_j)=$$

$$=\sum_{h=1}^{N}a_h^2P(x_i=a_h)-\sum_{h,k=1}^{N}a_ha_kP(x_i=a_h,\ x_j=a_k), \qquad \text{(XIV.4,4)}$$

deoarece

$$M(x_i^2)=\sum_{h=1}^{N} a_h^2 P(x_i=a_h) \text{ şi } M(x_i x_j)=\sum_{\substack{h,k=1\\ h\neq k}}^{N} a_h a_k P(x_i=a_h,\ x_j=a_k).$$

Tot din cele stabilite relativ la distribuţia hipergeometrică (cap. IX, § 3), se ştie că

$$P(x_i=a_h)=\frac{1}{N} \qquad (i=1,\ 2,\ \dots,\ n)\ (h=1,\ 2,\ \dots,\ N)$$

şi

$$P(x_i=a_h,\ x_j=a_k)=\frac{2}{N(N-1)} \qquad (i,\ \underset{i\neq j}{j=1,\ 2,\ \dots,\ n})\ (h,\ \underset{h\neq k}{k=1,\ 2,\ \dots,\ N})$$

Introducînd valorile acestor probabilităţi în (XIV.4,4), deducem

$$M\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2\right]=\frac{1}{N}\sum_{h=1}^{N}a_h^2-\frac{2}{N(N-1)}\sum_{\substack{h,k=1\\ h\neq k}}^{N}a_h a_k \quad \text{(XIV.4,5)}$$

sau, ţinînd seama de identitatea (XIV.4,3), obţinem

$$M\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2\right]=\frac{1}{N-1}\sum_{i=1}^{N}(x_i-\bar{x})^2=s^2,$$

ceea ce am vrut să demonstrăm.

Să observăm că, dacă unităţile selecţionate ale sondajului se pun înapoi în colectivitate, cu alte cuvinte variabilele de selecţie

$$x_1,\ x_2,\ \dots,\ x_n$$

sînt independente între ele, există relaţia

$$M\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2\right]=\frac{1}{N}\sum_{h=1}^{N}(a_h-a)^2=\sigma^2.$$

Aceasta se ştie, de altfel, din studiul momentelor centrate de selecţie (cap. XVI, § 3), unde am arătat că expresia

$$\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

reprezintă o estimaţie absolut corectă a momentului centrat teoretic σ^2.

2. Să considerăm, acum, cazul cînd fiecare unitate a colectivității A prezintă două caracteristici, adică unitatea de rangul j are coordonatele (a_j, b_j) sau, ceea ce este tot una, unitatea de rangul j este vectorul aleator

$$l_j(a_j, b_j) \qquad (j=1, 2, \ldots, N).$$

Efectuînd sondajul simplu de mărime n

$$x_1(u_1, v_1), \quad x_2(u_2, v_2), \ldots, x_n(u_n, v_n),$$

unde u_i, v_i sînt caracteristicile vectorului de selecție x_i, să introducem următoarele notații

$$\bar{u}=\frac{1}{n}\sum_{i=1}^{n} u_i, \quad \bar{v}=\frac{1}{n}\sum_{i=1}^{n} v_i,$$

$$\sigma_a^2=\frac{1}{N}\sum_{i=1}^{N}(a_i-a)^2, \quad \sigma_b^2=\frac{1}{N}\sum_{i=1}^{N}(b_i-b)^2,$$

$$\sigma_{ab}=\frac{1}{N}\sum_{i=1}^{N}(a_i-a)(b_i-b), \qquad \text{(XIV.4,6)}$$

$$s_a^2=\frac{1}{N-1}\sum_{i=1}^{N}(a_i-a)^2, \quad s_b^2=\frac{1}{N-1}\sum_{i=1}^{N}(b_i-b)^2,$$

$$s_{ab}=\frac{1}{N-1}\sum_{i=1}^{N}(a_i-a)(b_i-b).$$

Teoremă. *Sondajele simple de unități cu două caracteristici dau relațiile*

$$M(\bar{u})=a, \; M(\bar{v})=b,$$

$$M(\bar{u}-a)^2=(1-f)\frac{s_a^2}{n},$$

$$M(\bar{v}-b)^2=(1-f)\frac{s_b^2}{n}, \qquad \text{(XIV.4,7)}$$

$$M(\bar{u}-a)(\bar{v}-b)=(1-f)\frac{s_{ab}}{n},$$

unde $f=\frac{n}{N}$ $(n<N)$.

Primele patru relaţii se demonstrează în acelaşi mod ca în cazul sondajelor simple de unităţi cu o singură caracteristică. Rămîne să demonstrăm ultima relaţie.

Conform primelor două relaţii din (XIV.4,6), avem

$$(\overline{u}-a)(\overline{v}-b)=\frac{1}{n^2}\sum_{i,\,j=1}^{n}(u_i-a)(v_j-b).$$

De aci deducem

$$M(\overline{u}-a)(\overline{v}-b)=\frac{1}{n^2}\sum_{i,\,j=1}^{n}M(u_i-a)(v_j-b)=$$

$$=\frac{1}{n^2}\sum_{i=1}^{n}M(u_i-a)(v_i-b)+\frac{1}{n^2}\sum_{i,\,j=1}^{n}M(u_i-a)(v_j-b)=$$

$$=\frac{1}{n}\sum_{h=1}^{N}(a_h-a)(b_h-b)\,P(x_i=l_h)+$$

$$+\frac{n-1}{n}\sum_{\substack{h,\,k=1\\ h\neq k}}^{n}(a_h-a)(b_k-b)\,P(x_i=l_h,\ x_j=l_k), \qquad \text{(XIV.4,8)}$$

avînd în vedere că, potrivit definiţiei lor, $M(u_i-a)(v_i-b)=$

$$=\sum_{h=1}^{N}(a_h-a)(b_h-b)\,P(x_i=l_h) \text{ şi } M(u_i-a)(v_j-b)=$$

$$=\sum_{\substack{h,\,k=1\\ h\neq k}}^{n}(a_h-a)(b_k-b)\,P(x_i=l_h,\ x_j=l_k).$$

Trebuie, deci, să calculăm şi probabilităţile care intervin în formulele de mai sus. Astfel, probabilitatea ca u_i să ia valoarea a_h şi v_i să ia valoarea b_h, adică probabilitatea ca vectorul $x_i(u_i, v_i)$ să ia valoarea vectorială $l_h(a_h, b_h)$ este

$$P(x_i=l_h)=\frac{1}{N}.$$

In acelaşi mod avem

$$P(x_i=l_h,\ x_j=l_k)=\frac{1}{N(N-1)}\cdot$$
$$i\neq j,\quad h\neq k.$$

Probabilităţile

$$P(x_i=l_h,\ x_i=l_k)$$
$$h\neq k$$

şi

$$P(x_i=l_h,\ x_j=l_h)$$
$$i\neq j$$

sînt egale cu zero, deoarece prima reprezintă probabilitatea ca vectorul $x_i(u_i,\ v_i)$ să ia, în acelaşi timp, valorile vectoriale $l_h(a_h,\ b_h)$ şi $l_k(a_k,\ b_k)$, ceea ce este o imposibilitate. A doua probabilitate reprezintă probabilitatea ca vectorii diferiţi $x_i(u_i,\ v_i)$ şi $x_j(u_j,\ v_j)$ să ia aceeaşi valoare vectorială $l_h(a_h,\ b_h)$, ceea ce, de asemenea, este imposibil cunoscînd că sondajul se face fără a restitui unitatea extrasă.

Cu acestea, relaţia (XIV.4,8) devine

$$M(\bar{u}-a)(\bar{v}-b)=\frac{1}{nN}\sum_{h=1}^{N}(a_h-a)(b_h-b)+$$

$$+\frac{n-1}{nN(N-1)}\sum_{\substack{h,k=1\\h\neq k}}^{N}(a_k-a)(b_k-b)=\frac{1}{nN}\sum_{h=1}^{N}(a_h-a)(b_h-b)+$$

$$+\frac{n-1}{nN(N-1)}\sum_{h=1}^{N}(a_h-a)\sum_{\substack{k=1\\k\neq h}}^{N}(b_k-b). \qquad \text{(XIV.4,9)}$$

Insă, din egalitatea evidentă

$$\sum_{k=1}^{N}(b_k-b)=0,$$

deducem

$$\sum_{\substack{k=1\\k\neq h}}^{N}(b_k-b)=-(b_h-b).$$

In sfîrşit, înlocuind în relaţia (XIV.4,9), obţinem

$$M(\bar{u}-a)(\bar{v}-b)=\frac{1}{nN}\sum_{h=1}^{N}(a_h-a)(b_h-b)-\frac{n-1}{nN(N-1)}\sum_{h=1}^{N}(a_h-a)(b_h-b)=$$

$$=\frac{N-n}{N}\cdot\frac{1}{n(N-1)}\sum_{h=1}^{N}(a_h-a)(b_h-b)=(1-f)\frac{s_{ab}}{n}, \qquad \text{(XIV.4,10)}$$

relaţie pe care ne-am propus să o demonstrăm.

Mai general, se poate arăta, în acelaşi mod, că formulele demonstrate pentru selecţia cu variabile independente se păstrează şi în cazul selecţiei cu variabile dependente şi anume

$$M(\bar{u}-a)^{2m}=O\left(\frac{1}{n^m}\right),$$

$$M(\bar{v}-b)^{2m}=O\left(\frac{1}{n^m}\right),$$

$$M(\bar{u}-a)^m(\bar{v}-b)^m=O\left(\frac{1}{n^m}\right),$$

unde $O\left(\frac{1}{n^m}\right)$ este o funcţie asimptotică de ordinul lui $\frac{1}{n^m}$.

Teoremă. *Coeficientul de corelaţie $\rho_{\bar{u}\bar{v}}$ al variabilelor $\bar{u}$ şi $\bar{v}$ este independent de mărimea sondajului şi are ca expresie*

$$\rho_{\bar{u}\bar{v}}=\frac{s_{ab}}{s_a s_b}.$$

In baza definiţiei sale, coeficientul de corelaţie $\rho_{\bar{u}\bar{v}}$ se scrie

$$\rho_{\bar{u}\bar{v}}=\frac{M(\bar{u}-a)(\bar{v}-b)}{\sqrt{M(\bar{u}-a)}\cdot\sqrt{M(\bar{v}-b)}}$$

Scanned with OKEN Scanner

sau, înlocuind valorile medii care intră în această formulă prin valorile lor date de relaţiile (XIV.4,7), rezultă

$$\rho_{\bar{u}\bar{v}}=\frac{(1-f)\frac{s_{ab}}{n}}{\sqrt{(1-f)\frac{s_a^2}{n}}\cdot\sqrt{(1-f)\frac{s_b^2}{n}}}=\frac{s_{ab}}{s_a\cdot s_b},$$

deci expresia căutată care, după cum se observă, este independentă de n, mărimea sondajului efectuat.

3. In aplicaţiile practice ale statisticii prezintă un deosebit interes estimarea raportului

$$r=\frac{\bar{u}}{\bar{v}}$$

dintre variabilele $\bar{u}=\frac{1}{n}\sum_{i=1}^{n}u_i$ şi $\bar{v}=\frac{1}{n}\sum_{i=1}^{n}v_i$, unde u_i şi v_i sînt componentele vectorului de selecţie $x_i\,(i=1,\ 2,\ \dots,\ n)$.

Prin estimarea acestui raport se pot rezolva unele probleme legate de bugetele familiilor unei populaţii oarecare ca, de pildă, stabilirea raportului dintre cheltuielile totale făcute pentru hrană şi venitul total al familiei, sau stabilirea raportului dintre totalul cheltuielilor cu scop cultural şi distractiv, faţă de venitul total al familiei etc.

Problema estimării unui astfel de raport pe baza unui sondaj simplu cu două caracteristici, de mărime n,

$$x_1(u_1,\ v_1),\ x_2(u_2,\ v_2),\ \dots,\ x_n(u_n,\ v_n)$$

îşi găsesc rezolvarea completă în următoarea teoremă.

Teoremă. *Intr-un sondaj simplu, ale cărui unităţi prezintă două caracteristici, raportul* $\frac{\bar{u}}{\bar{v}}$ *este o estimaţie a raportului* $\frac{a}{b}$ *dintre mediile* $a=\frac{1}{N}\sum_{i=1}^{N}a_i$, *respectiv* $b=\frac{1}{N}\sum_{j=1}^{N}b_j$ *ale caracteristicilor colectivităţii totale, dacă* $a_i\geqq 0$, $b_j>0$ $(i,\ j=1,\ 2,\ \dots,\ N)$, *caracterizată prin relaţiile*

$$M\left(\frac{\bar{u}}{\bar{v}}\right)=\frac{a}{b}+O\left(\frac{1}{n}\right) \qquad \text{(XIV.4,11)}$$

şi

$$D^2\left(\frac{\overline{u}}{\overline{v}}\right)=(1-f)\frac{a^2}{b^2n}\left(\frac{s_a^2}{a^2}-2\frac{s_{ab}}{ab}+\frac{s_b^2}{b^2}\right)+O\left(\frac{1}{n}\right), \qquad \text{(XIV.4,12)}$$

unde $O\left(\frac{1}{n}\right)$ *reprezintă o funcţie de ordinul asimptotic al lui* $\frac{1}{n}$, *iar* $f=\frac{n}{N}$ $(n<N)$.

Presupunînd existenţa derivatelor parţiale de primele două ordine în raport cu $\overline{u}$ şi $\overline{v}$ ale funcţiei $f(\overline{u}, \overline{v})$, putem aplica dezvoltarea după formula lui Taylor în jurul punctului $(\overline{u}=a, \overline{v}=b)$

$$f(\overline{u}, \overline{v})=f(a, b)+(\overline{u}-a)f'_{\overline{u}}(a, b)+(\overline{v}-b)f'_{\overline{v}}(a, b)+$$

$$+\frac{1}{2}[(\overline{u}-a)^2 f''_{\overline{u}\,\overline{u}}(\xi, \eta)+2(\overline{u}-a)(\overline{v}-b)f''_{\overline{u}\,\overline{v}}(\xi, \eta)+$$

$$+(\overline{v}-b)^2 f''_{\overline{v}\,\overline{v}}(\xi, \eta)],$$

(ξ, η) fiind un punct intermediar între punctele (a, b) şi $(\overline{u}, \overline{v})$.

In cazul de faţă, $f(\overline{u}, \overline{v})=\frac{\overline{u}}{\overline{v}}$ şi formula devine, presupunînd $b\neq 0$,

$$\frac{\overline{u}}{\overline{v}}=\frac{a}{b}+\frac{1}{b}(\overline{u}-a)-\frac{a}{b^2}(\overline{v}-b)-\frac{1}{\eta^2}(\overline{u}-a)(\overline{v}-b)+$$

$$+\frac{\xi}{\eta^3}(\overline{v}-b)^2. \qquad \text{(XIV.4,13)}$$

Dacă notăm cu R' şi R'' respectiv mulţimea punctelor din spaţiul selecţiei, R_n, pentru care sînt îndeplinite inegalităţile

$$|\overline{u}-a|<\varepsilon, \quad |\overline{v}-b|<\varepsilon,$$

unde ε este un număr pozitiv oricît de mic, aceste inegalităţi vor defini în spaţiul R_n o regiune

$$\overline{R}=R'\cap R''.$$

Considerînd în spaţiul R_n punctul $M_{j_1 j_2 \ldots j_n}$, fiecare dintre indicii $j_1, j_2, \ldots, j_n$ luînd valori de la 1 la N, să determinăm probabili-

tatea ca acest punct să aparţină regiunii $\bar{R}$. Această probabilitate, în baza inegalităţii lui Boole, este

$$P(M_{j_1 j_2 \ldots j_n} \in \bar{R}) \geq P(M_{j_1 j_2 \ldots j_n} \in R') + P(M_{j_1 j_2 \ldots j_n} \in R'') - 1.$$

Observăm că, pentru determinarea probabilităţii $P(M_{j_1 j_2 \ldots j_n} \in \bar{R})$, trebuie să cunoaştem probabilităţile $P(M_{j_1 j_2 \ldots j_n} \in R')$ şi $P(M_{j_1 j_2 \ldots j_n} \in R'')$.

După inegalitatea lui Cebîşev şi avînd în vedere relaţiile (XIV.4,7), avem

$$P(M_{i_1 j_2 \ldots j_n} \in R') = P(|\bar{u} - a| < \varepsilon) \geq 1 - \frac{M(\bar{u}-a)^2}{\varepsilon^2} =$$

$$= 1 - \frac{(1-f)\, s_a^2}{n\varepsilon^2} = 1 - \frac{h}{n}, \quad \text{unde} \quad h = \frac{(1-f)\, s_a^2}{\varepsilon^2}.$$

In mod analog,

$$P(M_{j_1 j_2 \ldots j_n} \in R'') \geq 1 - \frac{k}{n},$$

unde

$$k = \frac{(1-f)\, s_b^2}{\varepsilon^2},$$

de unde urmează că

$$P(M_{j_1 j_2 \ldots j_n} \in \bar{R}) \geq 1 - \frac{l}{n}, \qquad \text{(XIV.4,14)}$$

dacă $l = h + k$ şi, evident,

$$P(M_{j_1 j_2 \ldots j_n} \in C\bar{R}) < \frac{l}{n}, \qquad \text{(XIV.4,15)}$$

$C\bar{R}$ fiind complementara regiunii $\bar{R}$.

Insemnînd, în general, prin

$$M_{\bar{R}}[F(\bar{u}, \bar{v})]$$

fracțiunea din valoarea medie a unei funcții $F(\bar{u}, \bar{v})$ pe regiunea $\bar{R}$ din spațiul R_n, vom scrie, în particular pentru dezvoltarea (XIV.4,13),

$$M_{\bar{R}}\left(\frac{\bar{u}}{\bar{v}}\right)=M_{\bar{R}}\left(\frac{a}{b}\right)+\frac{1}{b}M_{\bar{R}}(\bar{u}-a)-\frac{a}{b^2}M_{\bar{R}}(\bar{v}-b)-$$

$$-M_{\bar{R}}\times\left[\frac{1}{\eta^2}(\bar{u}-a)(\bar{v}-b)\right]+M_{\bar{R}}\left[\frac{\xi}{\eta^3}(\bar{v}-b)^2\right]. \qquad \text{(XIV.4,16)}$$

Urmează, deci, să calculăm fiecare dintre fracțiunile valorilor medii pe regiunea $\bar{R}$, care intră în această dezvoltare, atît în membrul întîi cît și în membrul al doilea.

In primul rînd să observăm, conform ipotezei noastre ($a_i \geqq 0$, $b_j > 0$; $i, j = 1, 2, \ldots, N$), că mediile $\bar{u}$ și $\bar{v}$ sînt mărginite superior, adică satisfac inegalitățile

$$\bar{u} < k', \quad \bar{v} < k''. \qquad \text{(XIV.4,17)}$$

In al doilea rînd observăm, în baza aceleiași ipoteze, că raportul $\frac{\bar{u}}{\bar{v}}$ este de asemenea mărginit superior în tot spațiul R_n și deci, implicit, în subspațiul $\bar{R} \in R_n$, deoarece avem

$$\frac{\bar{u}}{\bar{v}} < \frac{k'}{b_m} = k''', \qquad \text{(XIV.4,18)}$$

unde b_m reprezintă cea mai mică valoare dintre toate valorile $b_j > 0$ ($j = 1, 2, \ldots, N$), iar k''' este bineînțeles o constantă.

In sfîrșit, mai observăm că funcțiile $\frac{1}{\eta^2}$ și $\frac{\xi}{\eta^3}$ sînt continue în vecinătatea punctului ($\xi = a$, $\eta = b$), întrucît, după ipoteza de mai sus deducem

$$a = \frac{1}{N}\sum_{i=1}^{N} a_i \geqq 0, \quad b = \frac{1}{N}\sum_{j=1}^{N} b_j > 0.$$

Atunci, în virtutea continuității, putem alege un ε atît de mic, încît avînd

$$|\bar{u}-a| < \varepsilon, \quad |\bar{v}-b| < \varepsilon,$$

să existe un număr finit L, care să mărginească superior funcțiile $\frac{1}{\eta^2}$ și $\frac{\xi}{\eta^3}$

$$\frac{1}{\eta^2} < L, \quad \frac{|\xi|}{|\eta^3|} < L. \qquad \text{(XIV.4,19)}$$

Scanned with OKEN Scanner

In urma acestor consideraţii, fracţiunile din valorile medii ale dezvoltării (XIV.4,16) se calculează uşor. În adevăr, din primul membru al acestei dezvoltări deducem

$$M_{\overline{R}}\left(\frac{\overline{u}}{\overline{v}}\right)=M\left(\frac{\overline{u}}{\overline{v}}\right)-M_{C\overline{R}}\left(\frac{\overline{u}}{\overline{v}}\right).$$

Pe de altă parte

$$\left|M_{C\overline{R}}\left(\frac{\overline{u}}{\overline{v}}\right)\right| \leq M_{C\overline{R}}\left(\left|\frac{\overline{u}}{\overline{v}}\right|\right)<k''' P(M_{j_1 j_2 \ldots j_n} \in C\overline{R})<\frac{k'''l}{n},$$

ţinînd seamă de relaţiile (XIV.4,15) şi (XIV.4,16).
Rezultă, deci,

$$M_{\overline{R}}\left(\frac{\overline{u}}{\overline{v}}\right)=M\left(\frac{\overline{u}}{\overline{v}}\right)+O\left(\frac{1}{n}\right).$$

Procedînd la calcularea membrului al doilea al dezvoltării (XIV.4,16), obţinem, avînd în vedere relaţiile (XIV.4,14), (XIV.4,15) şi (XIV.4,17)

$$M_{\overline{R}}\left(\frac{a}{b}\right)=\frac{a}{b}\,P(M_{j_1 j_2 \ldots j_n} \in \overline{R}) \geq \frac{a}{b}\left(1-\frac{l}{n}\right)$$

sau

$$M_{\overline{R}}\left(\frac{a}{b}\right)=\frac{a}{b}+O\left(\frac{1}{n}\right).$$

Din egalităţile

$$M_{\overline{R}}(\overline{u}-a)+M_{C\overline{R}}(\overline{u}-a)=M(\overline{u}-a)=0$$

deducem

$$M_{\overline{R}}(\overline{u}-a)=-M_{C\overline{R}}(\overline{u}-a).$$

Cum, însă, din relaţiile de mai sus avem

$$|M_{C\overline{R}}(\overline{u}-a)| \leq M_{C\overline{R}}(\overline{u}-a)<k' P(M_{j_1 j_2 \ldots j_n} \in C\overline{R})<\frac{k'l}{n},$$

rezultă

$$\frac{1}{b}M_{\overline{R}}(\overline{u}-a)=O\left(\frac{1}{n}\right).$$

In acelaşi fel găsim

$$\frac{a}{b^2}M_{\overline{R}}(\overline{v}-b)=O\left(\frac{1}{n}\right).$$

Scanned with OKEN Scanner

Ultimele două fracţiuni ale valorilor medii din membrul al doilea al dezvoltării (XIV.4,16) se calculează aplicînd pe de o parte inegalitatea lui Schwarz şi folosind, pe de altă parte, relaţiile (XIV.4,10) şi (XIV.4,19). Astfel, obţinem

$$M_{\overline{R}}\left[\frac{1}{\eta^2}(\overline{u}-a)(\overline{v}-b)\right] \leq \sqrt{M_{\overline{R}}\left(\frac{1}{\eta^4}\right)} \cdot \sqrt{M_{\overline{R}}(\overline{u}-a)^2(\overline{v}-b)^2}.$$

Deoarece, potrivit relaţiilor amintite, avem

$$M_{\overline{R}}\left(\frac{1}{\eta^4}\right) \geqq L^2 \cdot P(M_{j_1 j_2 \ldots j_n}) < L^2$$

şi

$$M_{\overline{R}}(\overline{u}-a)^2(\overline{v}-b)^2 = O\left(\frac{1}{n^2}\right),$$

rezultă

$$M_{\overline{R}}\left[\frac{1}{\eta^2}(\overline{u}-a)(\overline{v}-b)\right] = O\left(\frac{1}{n}\right).$$

Fără nici o deosebire, se demonstrează că

$$M_{\overline{R}}\left[\frac{\xi}{\eta^3}(\overline{v}-b)^2\right] = O\left(\frac{1}{n}\right).$$

După cele stabilite mai sus, dezvoltarea (XIV.4,16) devine

$$M\left(\frac{\overline{u}}{\overline{v}}\right) = \frac{a}{b} + O\left(\frac{1}{n}\right).$$

Am demonstrat astfel relaţia (XIV.4,11), urmînd să demonstrăm şi relaţia (XIV.4,12).

Plecînd de la (XIV.4,13) avem

$$\frac{\overline{u}}{\overline{v}} - \frac{a}{b} = \left[\frac{1}{b}(\overline{u}-a) - \frac{a}{b^2}(\overline{v}-b)\right] -$$

$$- \left[\frac{1}{\eta^2}(\overline{u}-a)(\overline{v}-b) - \frac{\xi}{\eta^3}(\overline{v}-b)^2\right].$$

Ridicînd la pătrat, obţinem

$$\left(\frac{\overline{u}}{\overline{v}} - \frac{a}{b}\right)^2 = \frac{1}{b^2}(\overline{u}-a)^2 - \frac{2a}{b^3}(\overline{u}-a)(\overline{v}-b) + \frac{a^2}{b^4}(\overline{v}-b)^2 + R,$$

unde

$$R=-2\left[\frac{1}{b}(\bar{u}-a)-\frac{a}{b^2}(\bar{v}-b)\right]\left[\frac{1}{\eta^2}(\bar{u}-a)(\bar{v}-b)-\frac{\xi}{\eta^3}(\bar{v}-b)^2\right]+$$

$$+\left[\frac{1}{\eta^2}(\bar{u}-a)(\bar{v}-b)-\frac{\xi}{\eta^3}(\bar{v}-b)^2\right]^2.$$

Dispersia raportului $r=\frac{\bar{u}}{\bar{v}}$ va fi deci

$$D^2\left(\frac{\bar{u}}{\bar{v}}\right)=\frac{1}{b^2}M(\bar{u}-a)^2-\frac{2a}{b^3}M(\bar{u}-a)(\bar{v}-b)+\frac{a^2}{b^4}M(\bar{v}-b)^2+M(R)=$$

$$=\frac{1}{b^2}(1-f)\frac{s_a^2}{n}-\frac{2a}{b^3}(1-f)\frac{s_{ab}}{n}+\frac{a^2}{b^4}(1-f)\frac{s_b^2}{n}+M(R)=$$

$$=(1-f)\frac{a^2}{b^2n}\left(\frac{s_a^2}{a^2}-2\frac{s_{ab}}{ab}+\frac{s_b^2}{b^2}\right)+M(R), \qquad \text{(XIV.4,20)}$$

substituind mediile $M(\bar{u}-a)^2$, $M(\bar{u}-a)(\bar{v}-b)$ şi $M(\bar{v}-b)^2$ prin valorile lor date de relaţiile (XIV.4,7).

Calcule făcute după acelaşi procedeu ca mai sus şi pe care nu le mai reproducem aici ne arată că

$$M(R)=O\left(n^{-\frac{3}{2}}\right)$$

şi prin urmare dispersia raportului $r=\frac{\bar{u}}{\bar{v}}$ are ca expresie

$$D^2\left(\frac{\bar{u}}{\bar{v}}\right)=(1-f)\frac{a^2}{b^2n}\left(\frac{s_a^2}{a^2}-2\frac{s_{ab}}{ab}+\frac{s_b^2}{b^2}\right)+O\left(n^{-\frac{3}{2}}\right),$$

adică relaţia (XIV.4,12) din teoremă.

După cum se constată uşor, există convergenţa

$$D^2\left(\frac{\bar{u}}{\bar{v}}\right)\to 0$$

pentru valorile lui n suficient de mari, ceea ce ne spune că raportul $\frac{\bar{u}}{\bar{v}}$ este o estimaţie corectă a raportului $\frac{a}{b}$ dintre mediile caracteristicilor a_i respectiv b_j $(i, j=1, 2, \ldots, N)$ ale colectivităţii totale. Cu aceasta teorema este complet demonstrată.

Scanned with OKEN Scanner

Ca încheiere a acestui paragraf, este necesar să facem unele observaţii. Astfel, cînd efectuăm un sondaj de volum n dintr-o colectivitate finită cu N elemente, volumul sondajului poate fi, de pildă, egal cu oricare dintre numerele

$$\frac{N}{100}, \frac{2N}{100}, \ldots, \frac{sN}{100}, \ldots, \frac{10N}{100}.$$

Toate formulele stabilite în paragraful de faţă prezintă interes prin faptul că ele ne dau gradul de aproximaţie în raport cu volumul sondajului care, după cum am văzut, este întotdeauna mai mic decît volumul colectivităţii totale.

§ 5. Sondajul stratificat simplu

1) Să considerăm colectivitatea A împărţită în k straturi

$$A_1, A_2, \ldots, A_k,$$

ale căror unităţi prezintă cîte o singură caracteristică a_{ij} $(i=1, 2, \ldots, N_j)$ $(j=1, 2, \ldots, N_j)$. Cu alte cuvinte, fiecăruia dintre cele k straturi ale colectivităţii totale îi corespund, respectiv, caracteristicile

$$\begin{aligned} &a_{11}, a_{21}, \ldots, a_{N_1 1}, \\ &a_{12}, a_{22}, \ldots, a_{N_2 2}, \\ &\quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \\ &a_{1k}, a_{2k}, \ldots, a_{N_k k}. \end{aligned}$$

Din fiecare dintre aceste straturi facem respectiv cîte

$$n_1, n_2, \ldots, n_k$$

extracţii la întîmplare, astfel ca

$$n_1+n_2+\ldots+n_k=n,$$

fără ca unitatea extrasă să fie înapoiată în colectivitatea totală. Prin urmare, din fiecare strat efectuăm cîte un sondaj (teoretic), obţinînd astfel sondaje ale căror unităţi sînt respectiv

$$\begin{aligned} &x_{11}, x_{21}, \ldots, x_{n_1 1}, \\ &x_{12}, x_{22}, \ldots, x_{n_2 2}, \\ &\quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \\ &x_{1k}, x_{2k}, \ldots, x_{n_k k}, \end{aligned}$$

unde, evident, x_{ij} $(i=1, 2, \ldots, n_j)$ $(j=1, 2, \ldots, k)$ sînt de considerat ca variabile aleatorii.

Introducînd notaţiile

$$a=\frac{1}{N}\sum_{j=1}^{k}\sum_{i=1}^{N_j}a_{ij}$$

şi

$$a_j=\frac{1}{N_j}\sum_{i=1}^{N_j}a_{ij},$$

rezultă că

$$a=\frac{1}{N}\sum_{j=1}^{k}N_j a_j. \qquad \text{(XIV.5,1)}$$

Deci mediana caracteristicilor colectivităţii totale, a, poate fi interpretată ca media ponderată a mediilor caracteristicilor straturilor, ponderile fiind respectiv egale cu

$$\frac{N_j}{N} \qquad (j=1, 2, \ldots, k).$$

In mod analog, notînd

$$\bar{x}=\frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_j}x_{ij},$$

$$x_j=\frac{1}{n_j}\sum_{i=1}^{n_j}x_{ij},$$

deducem

$$\bar{x}=\frac{1}{n}\sum_{j=1}^{k}n_j\bar{x}_j,$$

deci media valorilor sondajului total de mărime $\sum_{j=1}^{k}n_j=n$ este egală cu media ponderată a mediilor valorilor sondajelor cu ponderea respectiv egală cu

$$\frac{n_j}{n} \qquad (j=1, 2, \ldots, k).$$

Teoremă. *Intr-un sondaj stratificat simplu, variabila aleatorie*

$$y=\frac{1}{N}\sum_{j=1}^{k}N_j\bar{x}_j \qquad \text{(XIV.5,2)}$$

este o estimație absolut corectă a mediei caracteristicilor colectivității totale pentru care

$$M(y)=a,$$

$$D^2(y)=\sum_{j=1}^{k}\frac{N_j^2}{N^2}(1-f_j)\frac{s_j^2}{n_j} \qquad \text{(XIV.5,3)}$$

cu

$$s_j^2=\frac{1}{N_j-1}\sum_{i=1}^{N_j}(a_{ij}-a_j)^2 \quad \text{şi} \quad f_j=\frac{n_j}{N_j}.$$

In adevăr, în baza celor stabilite la studiul sondajului simplu, avem pentru stratul A_j relațiile

$$M(\bar{x}_j)=a_j$$

şi

$$D^2(\bar{x}_j)=(1-f_j)\frac{s_j^2}{n_j}.$$

Cunoscînd că variabilele $\bar{x}_j$ sînt independente între ele, deoarece extracțiile din fiecare strat se fac independent unele de altele, putem scrie

$$M(y)=\frac{1}{N}\sum_{j=1}^{k}N_jM(\bar{x}_j)=\frac{1}{N}\sum_{j=1}^{k}N_ja_j=a$$

şi

$$D^2(y)=\frac{1}{N^2}\left[\sum_{j=1}^{k}N_j(\bar{x}_j-a)\right]^2=\frac{1}{N^2}\sum_{j=1}^{k}N_j^2D^2(\bar{x}_j)=\sum_{j=1}^{k}\frac{N_j^2}{N^2}(1-f_j)\frac{s_j^2}{n_j},$$

avînd în vedere relațiile (XIV.5,1) şi (XIV.5,2).
Am stabilit, deci, relațiile (XIV.5,3) din teorema enunțată.
Dacă numerele $n_j\,(j=1,\ 2,\ \ldots,\ k)$ sînt suficient de mari, se vede că $D^2(y)$ este foarte mic, devenind egal chiar cu zero, cînd fiecare dintre numerele n_j atinge, respectiv, valoarea maximă $N_j\,(j=1,\ 2,\ \ldots,\ k)$ şi, prin urmare, numai în aceste condiții se poate

afirma că variabila aleatorie y este o estimaţie absolut corectă a mediei caracteristicilor colectivităţii totale.

Această observaţie are o importanţă din punct de vedere practic, pentru că nu putem obţine o estimaţie a mediei a, decît dacă numărul extracţiilor din fiecare strat este suficient de mare.

Teorema este astfel demonstrată.

Prin definiţie, se numeşte *sondaj stratificat reprezentativ* sau *proporţional*, un sondaj stratificat simplu pentru care sînt îndeplinite egalităţile

$$f_j = f \qquad (j=1, 2, \ldots, k),$$

de unde

$$f = \frac{n_1}{N_1} = \frac{n_2}{N_2} = \ldots = \frac{n_k}{N_k} = \frac{\sum_{j=1}^{k} n_j}{\sum_{j=1}^{k} N_j} = \frac{n}{N},$$

din care deducem

$$n_j = \frac{n}{N} N_j \qquad (j=1, 2, \ldots, k).$$

Atunci, dispersia funcţiei de estimaţie (XIV.5,2), în cazul sondajului stratificat reprezentativ, este

$$D^2(y) = \frac{1-f}{n} \sum_{j=1}^{k} \frac{N_j}{N} s_j^2, \qquad \text{(XIV.5,4)}$$

ceea ce se poate uşor constata, introducînd valoarea de mai sus a lui n_j în expresia dispersiei din relaţiile (XIV.5,3).

Din studiul sondajelor simple se ştie că un sondaj simplu este mai puţin eficient decît un sondaj simplu reprezentativ. In cazul sondajului stratificat simplu, această proprietate nu este îndeplinită din punct de vedere teoretic, adică se pot construi sondaje stratificate simple care să fie mai eficiente decît sondajele stratificate reprezentative. Se poate arăta că pentru aceasta este suficient să aibă loc inegalitatea

$$n_j \leq N_j - \frac{N-n}{n} \qquad (j=1, 2, \ldots, k)$$

şi toate valorile dispersiilor s_j să fie foarte mici.

Scanned with OKEN Scanner

Din punct de vedere practic, observăm că inegalitatea de mai sus nu este niciodată satisfăcută, deoarece ar trebui ca fiecare N_j să fie mai mic decît ordinul sutelor ($N_j \leq 100$), ceea ce nu se întîmplă în realitate, ţinînd seamă că metoda sondajelor nu se aplică decît atunci cînd numărul unităţilor colectivităţii totale este foarte mare.

Tot din punct de vedere practic, mai observăm că sondajul stratificat reprezentativ este mai eficient decît sondajul simplu, care s-ar obţine independent de straturile colectivităţii totale.

Avem, pe de o parte,

$$\sum_{j=1}^{k}\sum_{i=1}^{k}(a_{ij}-a)^2=\sum_{j=1}^{k}\sum_{i=1}^{k}[(a_{ij}-a)+(a_j-a)]^2=$$

$$=\sum_{j=1}^{k}\sum_{i=1}^{k}(a_{ij}-a_j)^2+2\sum_{j=1}^{k}\sum_{i=1}^{N_j}(a_{ij}-a_j)(a_j-a)+\sum_{j=1}^{k}(a_j-a)^2=$$

$$=\left[\sum_{j=1}^{k}\sum_{i=1}^{N_j}(a_{ij}-a_j)^2+2\sum_{j=1}^{k}(a_j-a)\left(\sum_{i=1}^{N_j}a_{ij}-N_ja_j\right)\right]+\sum_{j=1}^{k}N_j(a_j-a)^2=$$

$$=\sum_{j=1}^{k}\sum_{i=1}^{N_j}(a_{ij}-a_j)^2+\sum_{j=1}^{k}N_j(a_j-a)^2,$$

iar

$$s^2=\frac{1}{N-1}\sum_{j=1}^{k}\sum_{i=1}^{N_j}(a_{ij}-a)^2$$

şi

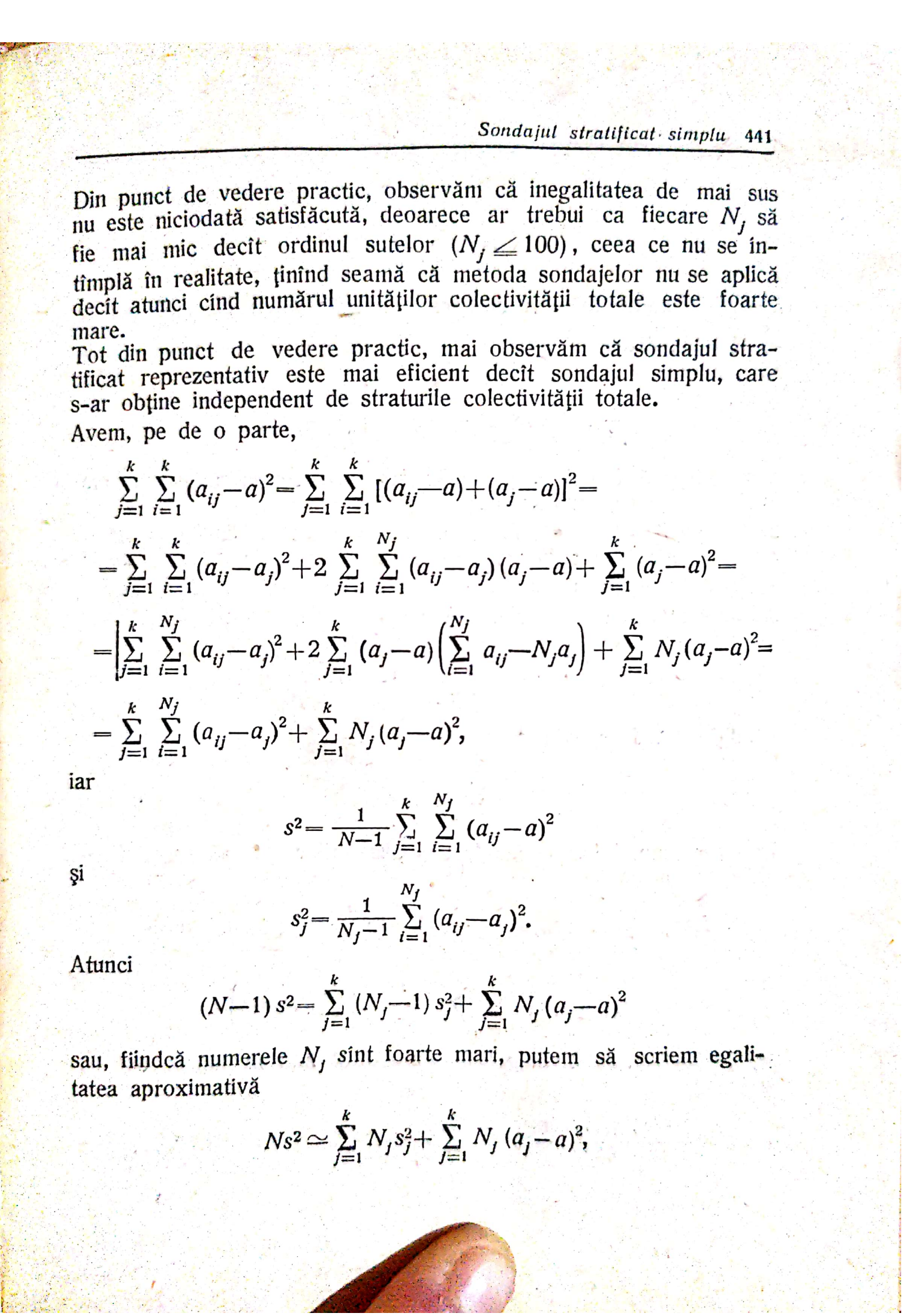
$$s_j^2=\frac{1}{N_j-1}\sum_{i=1}^{N_j}(a_{ij}-a_j)^2.$$

Atunci

$$(N-1)\,s^2=\sum_{j=1}^{k}(N_j-1)\,s_j^2+\sum_{j=1}^{k}N_j(a_j-a)^2$$

sau, fiindcă numerele N_j sînt foarte mari, putem să scriem egalitatea aproximativă

$$Ns^2\simeq\sum_{j=1}^{k}N_js_j^2+\sum_{j=1}^{k}N_j(a_j-a)^2,$$



deci

$$s^2 \geqq \sum_{j=1}^{k} \frac{N_j}{N} s_j^2 .$$

Inmulţind ambii membri ai acestei inegalităţi cu numărul pozitiv $\frac{1}{n}(1-f)$, deducem

$$\frac{1}{n}(1-f)\, s^2 \geqq \frac{1}{n}(1-f) \sum_{j=1}^{k} \frac{N_j}{N} s_j^2 ,$$

care exprimă faptul că sondajul stratificat reprezentativ este mai eficient decît sondajul simplu.

2. Să determinăm, acum, numărul de extracţii din fiecare strat, adică numărul unităţilor $n_j\,(j = 1,\ 2, \ldots, k)$ din fiecare sondaj astfel ca sondajul obţinut să fie cel mai eficient.

Aceasta revine a face minimă expresia

$$D^2(y) = \sum_{j=1}^{k} \frac{N_j^2}{N^2}(1-f_j)\frac{s_j^2}{n_j} ,$$

condiţionată de egalitatea

$$\sum_{j=1}^{k} n_j = n.$$

In acest scop, folosim metoda multiplicatorilor lui Lagrange. Considerînd funcţia

$$F(n_1, n_2, \ldots, n_k) = \sum_{j=1}^{k} \frac{N_j^2}{N^2}(1-f_j)\frac{s_j^2}{n_j} + \lambda\left(\sum_{j=1}^{k} n_j - n\right) =$$

$$= \sum_{j=1}^{k} \frac{N_j^2 s_j^2}{N^2 n_j} - \sum_{j=1}^{k} \frac{N s^2}{N^2} + \lambda\left(\sum_{j=1}^{k} n_j - n\right)$$

şi, anulînd derivatele ei parţiale de ordinul întîi în raport cu n_j, obţinem

$$\frac{\partial F(n_1, n_2, \ldots, n_k)}{\partial n_j} = -\frac{N_j^2 s_j^2}{N^2 n_j^2} + \lambda = 0, \qquad (j = 1,\ 2, \ldots, k),$$

ceea ce arată că λ este pozitiv. De aici

$$\frac{N_j s_j}{N n_j} = \sqrt{\lambda},$$

luînd pentru radical semnul plus, deoarece numerele N, N_j, s_j şi n_j sînt toate pozitive.
Ultima egalitate fiind echivalentă cu

$$\frac{\sum_{j=1}^{k} N_j s_j}{nN} = \sqrt{\lambda},$$

rezultă că

$$\frac{N_j s_j}{n_j} = \frac{\sum_{j=1}^{k} N_j s_j}{n},$$

de unde deducem

$$n_j = \frac{nN_j s_j}{\sum_{j=1}^{k} N_j s_j} \qquad (j = 1, 2, \ldots, k). \qquad \text{(XIV.5,5)}$$

Aceasta este, deci, expresia care determină valoarea numerelor n_j pentru care eficienţa sondajului este maximă. Un asemenea sondaj, al cărui număr de unităţi n este dat de formula (XIV.5,5), poartă numele de *sondaj stratificat optim*. După cum se observă, pentru acest sondaj numărul unităţilor sondajului dintr-un strat oarecare este direct proporţional cu numărul unităţilor din acel strat şi cu abaterea medie pătratrică a stratului respectiv.
Pentru a cunoaşte eficienţa sondajului stratificat optim, este necesar să calculăm dispersia funcţiei de estimaţie (XIV.5,2) corespunzătoare acestui sondaj. Ea se obţine înlocuind valoarea lui n_j din (XIV.5,5) în expresia lui $D^2(y)$ dată de (XIV.5,3). Astfel, găsim

$$D^2(y) = \sum_{j=1}^{k} \frac{N_j^2}{N^2}\left(1 - \frac{n_j}{N_j}\right)\frac{s_j^2}{n_j} = \sum_{j=1}^{k} \frac{N_j^2 s_j^2}{N_k^2}\left(\frac{1}{n_j} - \frac{1}{N_j}\right) =$$

$$= \sum_{j=1}^{k} \frac{N_j s_j^2}{N^2}\left(\frac{\sum_{j=1}^{k} N_j s_j}{n s_j} - 1\right) = \frac{1}{N^2}\left[\frac{\left(\sum_{j=1}^{k} N_j s_j\right)^2}{n} - \sum_{j=1}^{k} N_j s_j^2\right] \qquad \text{(XIV.5,6)}$$

Teoremă. *Sondajul stratificat optim este mai eficient decît sondajul stratificat reprezentativ.*

Trebuie, prin urmare, să demonstrăm că dispersia sondajului stratificat optim este mai mică decît dispersia sondajului stratificat reprezentativ. Atunci, calculînd diferenţa între dispersiile corespunzătoare celor două sondaje, date de relaţiile (XIV.5,4) respectiv (XIV.5,6), rezultă

$$\frac{1-f}{n}\sum_{j=1}^{k}\frac{N_j}{N}s_j^2-\frac{1}{N}\left[\frac{\left(\sum_{j=1}^{k}N_js_j\right)^2}{n}-\sum_{j=1}^{k}N_js_j^2\right]=$$

$$=\frac{1}{N}\left[\sum_{j=1}^{k}\frac{N_js_j^2}{N}-\left(\sum_{j=1}^{k}\frac{N_js_j}{N}\right)^2\right]=\frac{1}{n}\left[\sum_{j=1}^{k}\frac{N_j}{N}\left(s_j-\sum_{j=1}^{k}\frac{N_js_j}{N}\right)^2\right]>0.$$

Această inegalitate ne arată că dispersia sondajului stratificat reprezentativ este întotdeauna mai mare decit dispersia sondajului stratificat optim, ceea ce demonstrează teorema.

Tot de aici se mai vede că dispersia sondajului stratificat reprezentativ nu poate fi egală cu dispersia sondajului stratificat optim decît dacă toate numerele $s_j\,(j=1,\ 2,\ldots,k)$ sînt egale între ele, adică numai dacă dispersiile straturilor sînt aceleaşi.

3. Să examinăm cu acelaşi procedeu ca în cazul sondajelor dintr-o colectivitate infinită, cele două probleme legate de latura economică a sondajelor stratificate dintr-o colectivitate finită.

a) Cunoscînd cheltuielile totale C şi cheltuielile c_j pe o singură unitate din fiecare strat $A_j\,(j=1,\ 2,\ldots,k)$, să determinăm numărul unităţilor n_j corespunzătoare sondajului din fiecare strat, astfel ca să obţinem cea mai mare eficienţă a sondajului. Aceasta înseamnă să calculăm minimul dispersiei funcţiei de estimaţie dată de relaţia a doua din (XIV.5,3), adică

$$D^2(y)=\sum_{j=1}^{k}\frac{N_j^2}{N^2}(1-f_j)\frac{s_j^2}{n_j},$$

condiţionată de egalitatea

$$C=\sum_{j=1}^{k}n_jc_j.$$

Aplicînd metoda multiplicatorilor lui Lagrange, avem de considerat funcţia

$$F(n_1, n_2, \ldots, n_k) = \sum_{j=1}^{k} \frac{N_j^2}{N^2}(1-f_j)\frac{s_j^2}{n_j} + \lambda\left(\sum_{j=1}^{k} n_j c_j - C\right) =$$

$$= \sum_{j=1}^{k} \frac{N_j^2 s_j^2}{N^2 n_j} - \sum_{j=1}^{k} \frac{N_j s_j^2}{N^2} + \lambda\left(\sum_{j=1}^{k} n_j c_j - C\right).$$

Anulînd derivatele ei parţiale de ordinul întîi în raport cu n_j, obţinem

$$\frac{\partial F(n_1, n_2, \ldots, n_k)}{\partial n_j} = -\frac{N_j^2 s_j^2}{N^2 n_j^2} + \lambda c_j = 0 \qquad (j=1, 2, \ldots, k),$$

unde se vede de asemenea că $\lambda > 0$.
De aici rezultă

$$n_j = \frac{N_j s_j}{N\sqrt{\lambda}\cdot\sqrt{c_j}},$$

unde $\sqrt{\lambda}$ s-a luat cu semnul plus, deoarece n_j este, evident, un număr pozitiv.

Pe de altă parte, avînd în vedere relaţia $\sum_{j=1}^{k} n_j c_j = C$, deducem

$$C = \frac{\sum_{j=1}^{k} N_j s_j \sqrt{c_j}}{N\sqrt{\lambda}},$$

adică

$$\frac{1}{N\sqrt{\lambda}} = \frac{C}{\sum_{j=1}^{k} N_j s_j \sqrt{c_j}}.$$

Prin urmare, expresia lui n_j de mai sus devine

$$n_j = \frac{N_j s_j}{\sqrt{c_j}} \cdot \frac{C}{\sum_{j=1}^{k} N_j s_j \sqrt{c_j}} \qquad (j=1, 2, \ldots, k), \qquad \text{(XIV.5,7)}$$

care reprezintă numărul unităţilor sondajului din fiecare strat, în aşa fel încît să avem cea mai mare eficienţă a sondajului.

Scanned with OKEN Scanner

Din relaţia (XIV.5,7), prin însumare în raport cu indicele j de la 1 la k, deducem imediat mărimea sondajului, care este

$$n=C\sum_{j=1}^{k}\frac{N_j s_j}{\sqrt{c_j}\sum_{j=1}^{k}N_j s_j\sqrt{c_j}}\,.$$

In ceea ce priveşte mărimea dispersiei, care exprimă cea mai mare eficienţă a sondajului stratificat în condiţiile problemei date, se obţine astfel

$$D^2(y)=\sum_{j=1}^{k}\frac{N_j^2}{N^2}(1-f_j)\frac{s_j^2}{n_j}=\sum_{j=1}^{k}\frac{N_j^2 s_j^2}{N^2}\left(\frac{1}{n_j}-\frac{1}{N_j}\right)=$$

$$=\sum_{j=1}^{k}\frac{N_j s_j^2}{N^2}\left(\frac{\sqrt{c_j}\sum_{j=1}^{k}N_j s_j\sqrt{c_j}}{Cs_j}-1\right)=\frac{1}{N^2}\left[\frac{\sum_{j=1}^{k}N_j s\sqrt{c_j}}{C}-\sum_{j=1}^{k}N_j s_j^2\right]. \qquad \text{(XIV.5,8)}$$

In particular, dacă

$$c_1=c_2=\ldots=c_k=c$$

şi deci

$$C=c\sum_{j=1}^{k}n_j=cn,$$

regăsim, după cum ne aşteptam, expresiile (XIV.5,5) şi (XIV.5,6) ale lui n_j, respectiv $D^2(y)$ pentru cazul sondajului stratificat optim. In adevăr, relaţiile (XIV.5,7) şi (XIV.5,8) se pot atunci scrie

$$n_j=\frac{N_j s_j}{c}\cdot\frac{nc}{\sum_{j=1}^{k}N_j s_j}=\frac{nN_j s_j}{\sum_{j=1}^{k}N_j s_j}$$

şi

$$D^2(y)=\frac{1}{N^2}\left[\frac{c\left(\sum_{j=1}^{k}N_j s_j\right)^2}{nc}-\sum_{j=1}^{k}N_j s_j^2\right]=\frac{1}{N^2}\left[\frac{\left(\sum_{j=1}^{k}N_j s_j\right)^2}{n}-\sum_{j=1}^{k}N_j s_j^2\right].$$

b) Presupunînd că se cunoaşte eficienţa sondajului total, σ^2, se cere să se determine numărul unităţilor n_j ale sondajului din fiecare

strat A_j $(j=1, 2, \ldots, k)$ în aşa fel încît cheltuielile totale C să fie minime. Cu alte cuvinte, trebuie să găsim minimul expresiei

$$C=\sum_{j=1}^{k} n_j c_j,$$

condiţionată de egalitatea

$$\sigma^2=\sum_{j=1}^{k}\frac{N_j^2}{N^2}(1-f_j)\frac{s_j^2}{n_j}\cdot$$

Aplicînd, de asemenea, metoda multiplicatorilor lui Lagrange, trebuie să calculăm acum minimul funcţiei

$$F(n_1, n_2, \ldots, n_k)=\sum_{j=1}^{k} n_j c_j+\lambda\left[\sum_{j=1}^{k}\frac{N_j^2}{N^2}(1-f_j)\frac{s_j^2}{n_j}-\sigma^2\right]=$$

$$=\sum_{j=1}^{k} n_j c_j+\lambda\left[\sum_{j=1}^{k}\frac{N_j^2 s_j^2}{N^2 n_j}-\sum_{j=1}^{k}\frac{N_j s_j^2}{N^2}-\sigma^2\right].$$

Luînd derivatele parţiale de ordinul întîi în raport cu n_j ale funcţiei $F(n_1, n_2, \ldots, n_k)$ şi anulîndu-le, găsim

$$\frac{\partial F(n_1, n_2, \ldots, n_k)}{\partial n_j}=c_j-\lambda\frac{N_j^2 s_j^2}{N^2 n_j^2}=0, \qquad (j=1, 2, \ldots, k)$$

din care deducem

$$n_j=\frac{N_j s_j\sqrt{\lambda}}{N c_j},$$

semnul lui $\sqrt{\lambda}$ fiind evident pozitiv.
Atunci, expresia lui σ^2, care arată eficienţa sondajului, devine

$$\sigma^2=\sum_{j=1}^{k}\frac{N_j^2 s_j^2}{N^2 n_j}-\sum_{j=1}^{k}\frac{N_j s_j}{N^2}=\frac{1}{N\sqrt{\lambda}}\sum_{j=1}^{k} N_j s_j\sqrt{c_j}-\frac{1}{N^2}\sum_{j=1}^{k} N_j s_j^2.$$

Din această egalitate putem afla uşor pe $\sqrt{\lambda}$, care este egal cu

$$\sqrt{\lambda}=\frac{N\sum_{j=1}^{k} N_j s_j\sqrt{c_j}}{N^2\sigma^2+\sum_{j=1}^{k} N_j s_j^2}\cdot$$

În sfîrşit, introducînd valoarea lui $\sqrt{\lambda}$ în expresia lui n_j de mai sus, obţinem numărul unităţilor sondajului din fiecare strat, pentru care cheltuielile totale sînt minime, adică

$$n_j = \frac{N_j s_j}{\sqrt{c_j}} \cdot \frac{\sum_{j=1}^{k} N_j s_j \sqrt{c_j}}{N\sigma^2 + \sum_{j=1}^{k} N_j s_j^2} \qquad (j=1,\ 2,\ \ldots,\ k). \qquad \text{(XIV.5,9)}$$

Avînd valorile lui n_j, calculul cheltuielilor minime rezultă imediat şi anume

$$C = \sum_{j=1}^{k} n_j c_j = \sum_{j=1}^{k} N_j s_j \sqrt{c_j} \cdot \frac{\sum_{j=1}^{k} N_j s_j \sqrt{c_j}}{N\sigma^2 + \sum_{j=1}^{k} N_j s_j^2} = \frac{\left(\sum_{j=1}^{k} N_j s_j \sqrt{c_j}\right)^2}{N\sigma^2 + \sum_{j=1}^{k} N_j s_j^2}.$$

Observaţie. Numărul n_j al unităţilor sondajului din stratul A_j, calculat în condiţiile celor două probleme de minim examinate mai sus, este, după cum se vede din relaţiile (XIV.5,7) şi (XIV.5,9), direct proporţional cu numărul N_j al unităţilor din acel strat şi cu abaterea medie pătratică s_j a stratului respectiv şi invers proporţional cu rădăcina pătrată a cheltuielilor c_j pe o singură unitate din stratul corespunzător.

4. Să examinăm problema mai generală a sondajului stratificat simplu pentru cazul cînd unităţile fiecărui strat $A_j\,(j=1,\ 2,\ \ldots,\ k)$ prezintă cîte două caracteristici sau, altfel spus, unităţile colectivităţii totale A sînt vectorii aleatori

$$l_{ij}\,(a_{ij},\quad b_{ij}) \qquad (i=1,\ 2,\ \ldots,\ N_j) \quad (j=1,\ 2,\ \ldots,\ k).$$

Prin urmare, unităţile corespunzătoare fiecăruia dintre cele k straturi sînt respectiv

$$\begin{aligned} &l_{11},\ l_{21},\ \ldots,\ l_{N_1 1},\\ &l_{12},\ l_{22},\ \ldots,\ l_{N_2 2},\\ &\cdots\cdots\cdots\\ &l_{1k},\ l_{2k},\ \ldots,\ l_{Nkk}. \end{aligned}$$

Efectuînd cîte un sondaj de fiecare strat $A_j\,(j=1,\ 2,\ \ldots,\ k)$, unităţile sondajului stratificat simplu vor fi vectorii aleatori

$$x_{ij}\,(u_{ij},\quad v_{ij}). \qquad (i=1,\ 2,\ \ldots,\ n_j)\,(j=1,\ 2,\ \ldots,\ k).$$

Deci, din fiecare strat facem cîte un sondaj de mărime respectiv egală cu

$$\begin{aligned} &x_{11},\quad x_{21},\ \ldots,\ x_{n_1 1},\\ &x_{12},\quad x_{22},\ \ldots,\ x_{n_2 2},\\ &\cdot\quad\cdot\quad\cdot\quad\cdot\quad\cdot\quad\cdot\quad\cdot\\ &x_{1k},\quad x_{2k},\ \ldots,\ x_{n_k k}. \end{aligned}$$

Să introducem notaţia

$$l=\frac{1}{N}\sum_{j=1}^{k}N_j l_j,$$

adică vectorul mediu al colectivităţii totale, ale cărui componente sînt

$$a=\frac{1}{N}\sum_{j=1}^{k}N_j a_j,\quad b=\frac{1}{N}\sum_{j=1}^{k}N_j b_j, \qquad \text{(XIV.5,10)}$$

unde

$$a_j=\frac{1}{N_j}\sum_{i=1}^{N_j}a_{ij},\quad b_j=\frac{1}{N_j}\sum_{i=1}^{N_j}b_{ij}. \qquad \text{(XIV.5,11)}$$

In mod analog, pentru variabilele de selecţie introducem vectorul

$$\bar{x}=\frac{1}{N}\sum_{j=1}^{k}N_j\bar{x}_j,$$

ale cărui componente sînt

$$\bar{u}=\frac{1}{N}\sum_{j=1}^{k}N_j\bar{u}_j,\quad \bar{v}=\frac{1}{N}\sum_{j=1}^{k}N_j\bar{v}_j, \qquad \text{(XIV.5,12)}$$

unde

$$\bar{u}_j=\frac{1}{n_j}\sum_{i=1}^{n_j}u_{ij},\quad \bar{v}_j=\frac{1}{n_j}\sum_{i=1}^{n_j}v_{ij}. \qquad \text{(XIV.5,13)}$$

Scanned with OKEN Scanner

Să mai notăm, de asemenea,

$$s^2_{a_j a_j} = \frac{1}{N_j - 1} \sum_{i=1}^{N_j} (a_{ij} - a_j)^2, \quad s^2_{b_j b_j} = \frac{1}{N_j - 1} \sum_{i=1}^{N_j} (b_{ij} - b_j)^2,$$

$$s^2_{a_j b_j} = \frac{1}{N_j - 1} \sum_{i=1}^{N_j} (a_{ij} - a_j)(b_{ij} - b_j), \qquad \text{(XIV.5,14)}$$

$$\sigma^2_{a_j a_j} = \frac{1}{N_j} \sum_{i=1}^{N_j} (a_{ij} - a_j)^2, \quad \sigma^2_{b_j b_j} = \frac{1}{N_j} \sum_{i=1}^{N_j} (b_{ij} - b_j)^2,$$

$$\sigma^2_{a_j b_j} = \frac{1}{N_j} \sum_{i=1}^{N_j} (a_{ij} - a_j)(b_{ij} - b_j).$$

Teoremă. *Intr-un sondaj stratificat simplu, ale cărui unități prezintă cîte două caracteristici, componentele*

$$\bar{u}, \bar{v}$$

ale vectorului aleator $\bar{x}(\bar{u}, \bar{v})$ *sînt, respectiv, estimații absolut corecte pentru componentele* a *și* b *ale vectorului* $l\ (a, b)$.

Din relațiile (XIV.5,10) și (XIV.5,12) și aplicînd rezultatele stabilite la studiul sondajului simplu, deducem

$$M(\bar{u}) = \frac{1}{N} \sum_{j=1}^{k} N_j M(\bar{u}_j) = \frac{1}{N} \sum_{j=1}^{k} N_j a_j = a$$

și

$$M(\bar{v}) = \frac{1}{N} \sum_{j=1}^{k} N_j M(\bar{v}_j) = \frac{1}{N} \sum_{j=1}^{k} N_j b_j = b.$$

Vectorii $\bar{x}_j\ (\bar{u}_j, \bar{v}_j)$ fiind independenți între ei, deoarece extracțiile din fiecare strat nu depind unele de altele, putem scrie ca atare

$$D^2(\bar{u}) = \frac{1}{N^2} \sum_{j=1}^{k} N_j^2 D^2(\bar{u}_j)$$

și

$$D^2(\bar{v}) = \frac{1}{N^2} \sum_{j=1}^{k} N_j^2 D^2(\bar{v}_j).$$

Scanned with OKEN Scanner

Insă, în baza rezultatelor stabilite la studiul sondajului simplu, avem

$$D^2(\overline{u}_j) = (1-f_j)\frac{s^2_{a_j a_j}}{n_j}$$

şi

$$D^2(\overline{v}_j) = (1-f_j)\frac{s^2_{b_j b_j}}{n_j}.$$

Inlocuind aceste valori în expresiile de mai sus ale dispersiilor componentelor $\overline{u}$ şi $\overline{v}$, obţinem

$$D^2(\overline{u}) = \sum_{j=1}^{k} \frac{N_j^2}{N^2}(1-f_j)\frac{s^2_{a_j a_j}}{n_j}$$

şi

$$D^2(\overline{v}) = \sum_{j=1}^{k} \frac{N_j^2}{N^2}(1-f_j)\frac{s^2_{b_j b_j}}{n_j}.$$

Dacă numerele n_j cresc necontenit, rezultă că

$$D^2(\overline{u}) \to 0, \quad D^2(\overline{v}) \to 0.$$

Valoarea zero este atinsă de fiecare dintre aceste dispersii numai cînd numărul n_j ia cea mai mare valoare posibilă, care, după cum se ştie, este N_j.

Teorema este demonstrată.

Teoremă. *Intr-un sondaj stratificat simplu, ale cărui unităţi prezintă cîte două caracteristici, există relaţia*

$$M(\overline{u}-a)(\overline{v}-b) = \sum_{j=1}^{k} \frac{N_j^2}{N^2}(1-f_j)\frac{s^2_{a_j b_j}}{n_j}.$$

Din relaţiile (XIV.5,10) şi (XIV.5,12) urmează că

$$\overline{u}-a = \sum_{j=1}^{k} \frac{N_j}{N}(\overline{u}_j - a_j)$$

$$\overline{v}-a = \sum_{j=1}^{k} \frac{N_j}{N}(\overline{v}_j - b_j).$$

Prin urmare putem scrie

$$M(\bar{u}-a)(\bar{v}-b) = \sum_{j,h=1}^{k} \frac{N_j N_h}{N^2} M(\bar{u}_j - a_j)(\bar{v}_h - b_h) =$$

$$= \sum_{j=1}^{k} \frac{N_j^2}{N^2} M(\bar{u}_j - a_j)(\bar{v}_j - b_j) + \sum_{\substack{j,h=1 \\ j \neq h}}^{k} \frac{N_j N_h}{N^2} M(\bar{u}_j - a_j)(\bar{v}_h - b_h).$$

Observăm însă că, pentru $j \neq h$, vectorii $\bar{x}_j(\bar{u}_j, \bar{v}_j)$ şi $\bar{x}_h(\bar{u}_h, \bar{v}_h)$ sînt independenţi între ei. De aceea, suma

$$\sum_{\substack{j,h=1 \\ j \neq h}}^{k} \frac{N_j N_h}{N^2} M(\bar{u}_j - a_j)(\bar{v}_h - b_h) = \sum_{\substack{j,h=1 \\ j \neq h}}^{k} \frac{N_j N_h}{N^2} M(\bar{u}_j - a_j) M(\bar{v}_h - b_h) = 0.$$

Ţinînd apoi seama că, în urma celor stabilite la studiul sondajului simplu, avem

$$M(\bar{u}_j - a_j)(\bar{v}_j - b_j) = (1-f_j) \frac{s^2_{a_j b_j}}{n_j},$$

deducem relaţia căutată

$$M(\bar{u}-a)(\bar{v}-b) = \sum_{j=1}^{k} \frac{N_j^2}{N^2} (1-f_j) \frac{s^2_{a_j b_j}}{n_j},$$

ceea ce demonstrează teorema.

Teoremă. *Intr-un sondaj stratificat reprezentativ, coeficientul de corelaţie* $\rho_{\bar{u}\bar{v}}$ *al variabilelor* $\bar{u}$ *şi* $\bar{v}$ *este independent de mărimea sondajului.*

Potrivit definiţiei sale, coeficientul de corelaţie $\rho_{\bar{u}\bar{v}}$ este dat de expresia

$$\rho_{\bar{u}\bar{v}} = \frac{M(\bar{u}-a)(\bar{v}-b)}{\sqrt{M(\bar{u}-a)^2} \cdot \sqrt{M(\bar{v}-b)^2}},$$

care, după teorema precedentă, se mai poate scrie

$$\rho_{\bar{u}\bar{v}} = \frac{\sum_{j=1}^{k} \frac{N_j^2}{N^2} (1-f_j) \frac{s^2_{a_j b_j}}{n_j}}{\sqrt{\sum_{j=1}^{k} \frac{N_j^2}{N^2} (1-f_j) \frac{s^2_{a_j a_j}}{n_j}} \cdot \sqrt{\sum_{j=1}^{k} \frac{N_j^2}{N^2} (1-f_j) \frac{s^2_{b_j b_j}}{n_j}}}.$$

In cazul sondajului stratificat reprezentativ, avînd loc egalităţile

$$f_j = \frac{n_j}{N_j} = \frac{n}{N} = f,$$

rezultă

$$(1-f_j)\frac{1}{n_j} = \left(\frac{1}{f} - 1\right)\frac{1}{N_j}.$$

Atunci, expresia coeficientului de corelaţie devine

$$\rho_{\bar{u}\bar{v}} = \frac{\left(\frac{1}{f}-1\right)\frac{1}{N^2}\sum_{j=1}^{k} N_j s^2_{a_j b_j}}{\left(\frac{1}{f}-1\right)\sqrt{\sum_{j=1}^{k} N_j s^2_{a_j a_j}} \cdot \sqrt{\sum_{j=1}^{k} N_j s^2_{b_j b_j}}} =$$

$$= \frac{\sum_{j=1}^{k} N_j s^2_{a_j b_j}}{\sqrt{\sum_{j=1}^{k} N_j s^2_{a_j a_j}} \cdot \sqrt{\sum_{j=1}^{k} N_j s^2_{b_j b_j}}},$$

care, după cum se constată, este independentă de mărimea sondajului.

Observaţie. Această proprietate, stabilită şi pentru sondajul simplu cu unităţile avînd cîte două caracteristici, nu se păstrează pentru sondajul stratificat simplu avînd unităţile cu două caracteristici, decît numai în cazul sondajului stratificat reprezentativ.

5. Am arătat, în cazul sondajului simplu cu două caracteristici, că de multe ori ne interesează să cunoaştem raportul dintre mediile a şi b ale caracteristicilor colectivităţii totale. De aceea în continuare, ne vom ocupa de aceeaşi problemă, referitoare la sondajul stratificat simplu. Păstrînd aceleaşi notaţii de la punctul 4 al acestui paragraf, putem atunci enunţa teorema de mai jos.

Teoremă. *Intr-un sondaj stratificat simplu, raportul* $\frac{\bar{u}}{\bar{v}}$ *dintre componentele vectorului mediu aleator* $\bar{x}(\bar{u}, \bar{v})$ *al sondajului stratificat simplu reprezintă o estimaţie corectă pentru raportul* $\frac{a}{b}$ *dintre componentele vectorului mediu* $l(a, b)$ *al colectivităţii totale, avînd deci relaţiile*

$$M\left(\frac{\bar{u}}{\bar{v}}\right) = \frac{a}{b} + O\left(\frac{1}{n_l}\right)$$

şi

$$D^2\left(\frac{\overline{u}}{\overline{v}}\right)=\frac{a^2}{b^2}\left[\frac{1}{a^2}\sum_{j=1}^{k}\frac{N_j^2}{N^2}(1-f_j)\frac{s_{a_ja_j}^2}{n_j}+\frac{2}{ab}\sum_{j=1}^{k}\frac{N_j^2}{N^2}(1-f_j)\frac{s_{a_jb_j}^2}{n_j}+\right.$$

$$\left.+\frac{1}{b^2}\sum_{j=1}^{k}\frac{N_j^2}{N^2}(1-f_j)\frac{s_{b_jb_j}^2}{n_j}\right]+O\left(\frac{1}{n_i^{\frac{3}{2}}}\right),$$

unde notăm în general prin $O\left(\frac{1}{n_i^m}\right)$ *o funcţie asimptotică de ordinul lui* $\frac{1}{n_i^m}$, n_i *fiind cel mai mic dintre numerele* n_j $(j=1, 2, \ldots, k)$,

iar $f_j=\frac{n_j}{N_j}$ $(n_j<N_j)$.

Demonstraţia se face uşor, în baza teoremei de la punctul 3 al paragrafului precedent. In adevăr, s-a demonstrat acolo că, dacă $\overline{u}$ şi $\overline{v}$ sînt două variabi'e aleatorii pentru care avem relaţiile

$$M(\overline{u})=a, \quad M(\overline{v})=b,$$

iar momentele lor centrate de ordinul al doilea sînt funcţii $O\left(\frac{1}{n}\right)$, adică asimptotic de ordinul lui $\frac{1}{n}$, atunci avem, de asemenea relaţiile (XIV.4,11), respectiv (XIV.4,12) şi anume

$$M\left(\frac{\overline{u}}{\overline{v}}\right)=\frac{a}{b}+O\left(\frac{1}{n}\right)$$

şi

$$D^2\left(\frac{\overline{u}}{\overline{v}}\right)=\frac{a^2}{b^2}\left[\frac{1}{a^2}M(\overline{u}-a)^2+\frac{2}{ab}M(\overline{u}-a)(\overline{v}-b)+\right.$$

$$\left.+\frac{1}{b^2}M(\overline{v}-b)^2\right]+O\left(\frac{1}{n}\right).$$

Pe de altă parte, am mai demonstrat la punctul 4 din acest paragraf următoarele relaţii

$$M(\overline{u}-a)^2=\sum_{j=1}^{k}\frac{N_j^2}{N^2}(1-f_j)\frac{s_{a_ja_j}^2}{n_j},$$

$$M(\overline{u}-a)(\overline{v}-b)=\sum_{j=1}^{k}\frac{N_j^2}{N^2}(1-f_j)\frac{s_{a_jb_j}^2}{n_j}$$

şi

$$M(\bar{v}-b)^2=\sum_{j=1}^{k}\frac{N_j^2}{N^2}(1-f_j)\frac{s_{b_jb_j}^2}{n_j},$$

care exprimă faptul că momentele centrate de ordinul al doilea ale variabilelor aleatorii $\bar{u}$ şi $\bar{v}$, componentele vectorului mediu aleator $\bar{x}(\bar{u},\bar{v})$ al sondajului stratificat simplu, sînt funcţii asimptotice de ordinul lui $\frac{1}{n_i}$, n_i fiind cel mai mic număr dintre numerele n_j $(j=1,2,\ldots,k)$.

Atunci, ţinînd seama de aceste relaţii, precum şi de relaţiile (XIV.4,11) şi (XIV.4,12), teorema, în cazul nostru, rezultă imediat.

Capitolul XV

Sondaje cu două faze

§ 1. Sondajul cu două faze cu o singură caracteristică

După cum am arătat în § 3 al capitolului precedent, sondajul cu două faze se obţine alegînd în prima fază $h<k$ straturi la întîmplare dintr-o colectivitate A împărţită în k straturi

$$A_1, A_2, \ldots, A_k$$

şi apoi, în a doua fază, efectuînd din cele h straturi ieşite, extracţii în acelaşi mod ca la sondajul stratificat simplu. Colectivitatea totală şi componenţa straturilor admitem că este aceeaşi ca la sondajul stratificat simplu. Sondajul stratificat simplu este un caz particular al sondajului cu două faze, pentru $h=k$.

Aceleaşi probleme care se pun pentru sondajul stratificat simplu se pun şi pentru sondajul cu două faze, care se impune de preferinţă atunci cînd trebuie să se efectueze o statistică într-un timp mai scurt sau cînd suma totală a cheltuielilor este prea mare.

De exemplu, să presupunem că într-o ţară sînt 10 000 comune rurale. Pe baza unui sondaj, se cere evaluarea numărului de familii cu mai mult de cinci copii, din comunele rurale. Considerînd că fiecare comună rurală reprezintă un strat, putem să estimăm numă-

rul de familii pe care-l căutăm făcînd, de pildă, la întîmplare, cîte un sondaj de 10% din fiecare comună rurală. Procedînd astfel, aplicăm metoda sondajului stratificat simplu. Această metodă necesită însă un timp de lucru mai îndelungat, deoarece conduce la un număr de unităţi mai mare.

De aceea se poate proceda la un sondaj cu două faze, alegînd la întîmplare, în prima fază 100 comune rurale, de pildă, din totalul de 10 000 comune rurale. Apoi, în a doua fază, din fiecare dintre aceste 100 comune rurale, alegem la întîmplare, de exemplu, un procent de 10%. Se vede, deci, utilitatea sondajului cu două faze din punct de vedere economic.

Rezolvarea unei astfel de probleme îşi găseşte soluţionarea în teorema de mai jos. Mai înainte, însă, este necesar să dăm următoarele două leme, privind probabilităţile condiţionate.

Lema 1. *Dacă* ξ *şi* η *sînt două variabile aleatorii, avem relaţiile*

$$M(\eta)=M[M(\eta\,|\,\xi)] \tag{XV.1,1}$$

şi

$$D^2(\eta)=D^2[M(\eta\,|\,\xi)]+M[D^2(\eta\,|\,\xi)]. \tag{XV.1,2}$$

Fie atunci

$$F(x)=P(\xi\leqslant x)$$

funcţia de repartiţie a variabilei ξ,

$$G(y\,|\,x)=P(\eta\leqslant y\,|\,\xi=x)$$

funcţia de repartiţie a variabilei η, condiţionată de faptul că variabila $\xi=x$, şi

$$H(y)=P(\eta\leqslant y)$$

funcţia de repartiţie a variabilei η.

Notînd, respectiv, prin

$$X,\ Y$$

mulţimea evenimentelor variabilelor ξ şi η, putem scrie

$$dH(y)=\int_X dF(x)\,dG(y\,|\,x).$$

Pe de altă parte, în baza definiţiei mediei, avem

$$M(\eta)=\int_Y y\,dH(y),$$

sau, înlocuind pe $dH(y)$, obţinem

$$M(\eta)=\int_Y y\int_X dF(x)\,dG(y\mid x)=\int_X dF(x)\int_Y y\,dG(y\mid x)=M[M(\eta\mid\xi)],$$

adică relaţia (XV.1,1) din lema 1.

Rămîne să stabilim şi relaţia (XV.1,2). Pentru aceasta, avem de asemenea, în baza definiţiei mediei,

$$M(\eta^2)=\int_Y y^2 dH(y)=\int_Y y^2\int_X dF(x)\,dG(y\mid x)=$$

$$=\int_X dG(x)\int_Y y^2 dG(y\mid x)=M[M(\eta^2\mid\xi)].$$

Ţinînd seama de valorile obţinute pentru $M(\eta)$ şi $M(\eta^2)$, rezultă

$$D^2(\eta)=M(\eta^2)-[M(\eta)]^2=M[M(\eta^2\mid\xi)]-\{M[M(\eta\mid\xi)]\}^2=$$

$$=M\{[M(\eta^2\mid\xi)]-[M(\eta\mid\xi)^2]\}+M\{[M(\eta\mid\xi)]^2\}-\{M[M(\eta\mid\xi)]\}^2=$$

$$=M[D^2(\eta\mid\xi)]+D^2[M(\eta\mid\xi)],$$

care este relaţia (XV.1,2) din lema 1.

Am demonstrat astfel lema 1.

Lema 2. *Dacă* ξ, η *şi* ζ *sînt trei variabile aleatorii, iar*

$$C(uv)=M\{[u-M(u)][v-M(v)]\}$$

este covarianţa variabilelor aleatorii u *şi* v, *avem relaţia*

$$C(\eta,\ \xi)=M[C(\eta\zeta\mid\xi)]+C[M(\eta\mid\xi)\,M(\zeta\mid\xi)]. \qquad \text{(XV.1,3)}$$

Fie de asemenea

$$F(x)=P(\xi\leqslant x)$$

funcţia de répartiţie a variabiléi ξ,

$$G(y,z\mid x)=P(\eta\leqslant y,\ \zeta\leqslant z\mid\xi=x)$$

funcţia de repartiţie a variabilelor η şi ξ, condiţionate de faptul că variabila $\xi=x$ şi

$$H(y,\ z)=P(\eta\leqslant y,\ \zeta\leqslant z)$$

funcţia de repartiţie a variabilelor η şi ζ.

Dacă notăm respectiv prin

$$X,\ Y,\ Z$$

mulţimea tuturor evenimentelor variabilelor ξ, η şi ζ, putem scrie

$$dH(y, z) = \int_X dF(x)\, dG(y, z \mid x),$$

Insă, după definiţia mediei, avem

$$M(\eta\zeta) = \int_Y \int_Z yz\, dH(y, z) = \int_Y \int_Z yz \int_X dF(x)\, dG(y, z \mid x) =$$

$$= \int_X dF(x) \int_Y \int_Z yz\, dG(y, z \mid x) = \int_X dF(x)\, M(\eta\zeta \mid x) = M[M(\eta\zeta \mid \xi)],$$

înlocuind pe $dH(y, z)$ cu valoarea obţinută.
De unde rezultă relaţia (XV.1,3) din lema 2, adică

$$C(\eta\zeta) = M(\eta\zeta) - M(\eta)\, M(\zeta) = M[M(\eta\zeta \mid \xi)] - M[M(\eta \mid \xi)] \cdot M[M(\zeta \mid \xi)] =$$

$$= M[M(\eta\zeta \mid \xi) - M(\eta \mid \xi) \cdot M(\zeta \mid \xi)] +$$

$$+ M[M(\eta \mid \xi)\, M(\zeta \mid \xi)] - M[M(\eta \mid \xi)] \cdot M[M(\zeta \mid \xi)] =$$

$$= M[C(\eta\zeta \mid \xi)] + C[M(\eta \mid \xi) \cdot M(\zeta \mid \xi)],$$

folosind valoarea lui $M(\eta\zeta)$ de mai sus şi avînd în vedere relaţia (XV.1,1) din lema 1.

Teoremă. *Intr-un sondaj cu două faze, avînd o singură caracteristică, dacă notăm prin*

$$x_{ij} \qquad (i = 1, 2, \ldots, n_j)$$

variabilele aleatorii referitoare la unităţile secundare ale sondajului efectuat din stratul A_j $(j = 1, 2, \ldots, k)$ *şi prin*

$$Y_i = \frac{kN_j}{N} \cdot \frac{\sum_{i=1}^{n_j} x_{ij}}{n_j} \qquad \text{(XV.1,4)}$$

şi

$$Y = \frac{Y_{j_1} + Y_{j_2} + \ldots + Y_{j_h}}{h}, \qquad \text{(XV.1,5)}$$

atunci au loc relaţiile

$$M(Y) = a \qquad \text{(XV.1,6)}$$

şi

$$D^2(Y)=\frac{k(k-h)}{hN^2}\cdot\frac{\sum_{j=1}^{k}(\alpha_j-\alpha)^2}{k-1}+\frac{k}{hN^2}\cdot\sum_{j=1}^{k}N_j^2(1-f_j)\frac{f_j^2}{n_j},\qquad\text{(XV.1,7)}$$

unde

$$a=\frac{1}{N}\sum_{j=1}^{k}\sum_{i=1}^{N_j}a_{ij},\quad \alpha_j=\sum_{i=1}^{N_j}a_{ij},\quad \alpha=\frac{1}{k}\sum_{j=1}^{k}\alpha_k \text{ iar } f_j \text{ şi } f_j^2$$

au aceleaşi semnificaţii ca în cazul sondajului simplu cu o singură caracteristică.

1. Observăm, de la început, că în acest caz avem de efectuat două operaţii cu caracter aleator. Astfel, în prima operaţie, alegînd la întîmplare $h<k$ straturi

$$X_1,\ X_2,\ \ldots,\ X_h,$$

sîntem evident în prezenţa unui număr de h variabile aleatorii. Dacă am obţinut în prima operaţie straturile

$$X_1=A_{j_1},\ X_2=A_{j_2},\ \ldots,\ X_h=A_{j_h},$$

atunci, în operaţia a doua, făcînd respectiv la întîmplare

$$\eta_{ij_1},\ \eta_{ij_2},\ \ldots,\ \eta_{ij_h}$$

extracţii din fiecare strat ieşit, avem respectiv valorile

$$Y_{j_1},\ Y_{j_2},\ \ldots,\ Y_{j_h}$$

care sînt, de asemenea, nişte variabile aleatorii.
Considerînd media (XV.1,5)

$$Y=\frac{Y_{j_1}+Y_{j_2}+\ldots+Y_{j_h}}{h}$$

obţinem o nouă variabilă aleatorie, avînd în vedere, în acelaşi timp, cînd ne ocupăm de proprietăţile sale, că şi indicii

$$j_1,\ j_2,\ \ldots,\ j_h$$

sînt la rîndul lor variabile aleatorii.
Fie atunci

$$P(j_1,\ j_2,\ \ldots,\ j_h)$$

probabilitatea de a obține indicii $j_1, j_2, \ldots, j_h$, adică probabilitatea ca, din h extracții din cele k unități primare, să obținem respectiv straturile

$$A_{j_1}, A_{j_2}, \ldots, A_{j_h}.$$

Pe de o parte, în baza relației (XV.1,1) din lema 1, deducem

$$M(Y)=M[Y|(j_1, j_2, \ldots, j_h)], \qquad \text{(XV.1,8)}$$

înțelegînd prin

$$Y|(j_1\ j_2, \ldots, j_h)$$

variabila aleatorie Y, condiționată de faptul că s-au obținut straturile

$$A_{j_1}, A_{j_2}, \ldots, A_{j_h},$$

iar, pe de altă parte, rezultă în mod direct că

$$M[Y|(j_1, j_2, \ldots, j_h)]=\frac{1}{h}\sum_{r=1}^{h} M[Y_{j_r}|(j_1, j_2, \ldots, j_h)]. \qquad \text{(XV.1,9)}$$

Însă din relația (XV.1,4) și cunoscînd formulele stabilite la studiul sondajului simplu cu o singură caracteristică, avem

$$M(Y_j)=\frac{k}{N}N_j a_j.$$

De unde urmează că

$$M[Y|(j_1, j_2, \ldots, j_h)]=\frac{k}{hN}\sum_{r=1}^{h} N_{j_r} a_{j_r}. \qquad \text{(XV.1,9')}$$

Prin urmare $M[Y|(j_1, j_2, \ldots, j_h)]$ este variabila aleatorie, care cu probabilitatea $P(j_1, j_2, \ldots, j_h)$ ia valoarea dată de expresia (XV.1,9). Atunci relația (XV.1,8) se scrie

$$M(Y)=M[M(Y|(j_1, j_2, \ldots, j_h)]=$$

$$=\sum_{j_1, j_2 \ldots j_h} P(j_1, j_2, \ldots, j_h)\frac{k}{hN}\sum_{r=1}^{h} N_{j_r} a_{j_r}=$$

$$=\frac{1}{k}\sum_{j=1}^{k}\frac{kN_j a_j}{N}=\frac{1}{N}\sum_{j=1}^{k} N_j a_j=a,$$

deoarece, din studiul sondajului simplu cu o singură caracteristică, se ştie că, dacă

$$b_1,\ b_2,\ \ldots,\ b_k$$

sînt unităţile sondajului, valoarea mediei este

$$\sum_{j_1,\ j_2,\ \ldots,\ j_h} P(j_1,\ j_2,\ \ldots,\ j_h)\,\frac{1}{h}\sum_{r=1}^{h} b_{j_r} = \frac{1}{k}\sum_{j=1}^{k} b_j,$$

Am stabilit relaţia (XV.1,6) din teoremă.

Să stabilim şi relaţia (XV.1,7) din aceeaşi teoremă. Din (XV.1,2) deducem imediat

$$D^2(Y) = D^2[M(Y|(j_1,\ j_2,\ldots,\ j_h))] + M[D^2(Y|(j_1,\ j_2,\ldots,\ j_h))]. \quad \text{(XV.1,10)}$$

Urmează, deci, să calculăm fiecare dintre cei doi termeni ai membrului al doilea din această egalitate. Vom avea, astfel, penttu primul termen

$$D^2[M(Y|(j_1,\ j_2,\ \ldots,\ j_h))] =$$

$$= \sum_{j_1,\ j_2\ \ldots,\ j_h} P(j_1,\ j_2,\ldots j_h)\left(\frac{k}{hN}\sum_{r=1}^{h} N_{j_r}a_{j_r} - a\right)^2, \qquad \text{(XV.1,11)}$$

deoarece avem de calculat dispersia variabilei aleatorii

$$M[Y|(j_1,\ j_2,\ \ldots,\ j_h)]$$

care, după cum am văzut, ia cu probabilitatea $P(j_1,\ j_2,\ldots,\ j_h)$ valoarea dată de relaţia (XV.1,9′) şi are media egală cu a.

Însă, la determinarea dispersiei funcţiei de estimaţie, considerate la studiul sondajului simplu cu o singură caracteristică, am arătat că, dacă unităţile sondajului sînt aceleaşi ca mai sus, valoarea dispersiei este

$$\sum_{j_1,\ j_2,\ \ldots,\ j_h} P(j_1,\ j_2,\ \ldots,\ j_h)\left(\frac{1}{h}\sum_{r=1}^{h} b_{j_r} - b\right)^2 = \frac{k-h}{kh}\,\frac{\sum_{j=1}^{k}(b_j-b)^2}{k-1},$$

unde

$$b = \frac{1}{k}\sum_{j=1}^{k} b_j.$$

In cazul nostru, avînd

$$b_j = \frac{k}{N} N_j a_j,$$

relaţia (XV.1,11) devine

$$D^2[M(Y|(j_1, j_2, \ldots, j_h))] = \frac{k-h}{kh} \cdot \frac{\sum_{j=1}^{k} \left(\frac{k}{N} N_j a_j - a\right)^2}{k-1} =$$

$$= \frac{k-h}{hN^2} \cdot \frac{\sum_{j=1}^{k} (\alpha_j - \alpha)^2}{k-1}, \qquad \text{(XV.1,12)}$$

însemnînd

$$\alpha_j = \sum_{i=1}^{N_j} a_{ij} \quad \text{şi} \quad \alpha = \frac{1}{k} \sum_{j=1}^{k} \alpha_j.$$

Să trecem acum la determinarea termenului al doilea din egalitatea (XV.1,10), adică la calculul lui $M[D^2(Y|(j_1, j_2, \ldots, j_h))]$.

Din studiul sondajului stratificat simplu, cu o singură caracteristică, avem pentru indicii $j_1, j_2, \ldots, j_h$ fixaţi

$$D^2[Y|(j_1, j_2, \ldots, j_h)] = \frac{1}{h^2} \sum_{r=1}^{h} D^2(Y_{j_r}) =$$

$$= \frac{1}{h^2} \cdot \frac{k^2}{N^2} \sum_{r=1}^{h} N_{j_r}^2 D^2\left(\frac{\sum_{i=1}^{n_{j_r}} x_{ij_r}}{n_{j_r}}\right) =$$

$$= \frac{k^2}{h^2 N^2} \sum_{r=1}^{h} N_{j_r}^2 (1-f_j) \frac{f_{ir}^2}{n_{j_r}}. \qquad \text{(XV.1,13)}$$

Prin urmare, $D^2[Y|(j_1, j_2, \ldots, j_h)]$ este variabila aleatorie, care cu probabilitatea $P(j_1, j_2, \ldots, j_h)$ ia valoarea dată de (XV.1,13). Rezultă atunci că

$$M[D^2(Y|(j_1, j_2, \ldots, j_h))] =$$

$$= \sum_{j_1, j_2, \ldots, j_h} P(j_1, j_2, \ldots, j_h) \frac{1}{h} \left[\frac{k^2}{hN^2} \sum_{r=1}^{h} N_{j_r}^2 (1-f_{j_r}) \frac{f_{j_r}^2}{n_{j_r}}\right] =$$

$$= \frac{k^2}{h^2 N^2} \sum_{j=1}^{k} N_j^2 (1-f_j) \frac{f_j^2}{n_j}, \qquad \text{(XV.1,14)}$$

Scanned with OKEN Scanner

reamintind că, din studiul sondajului simplu cu o singură caracteristică, avem relaţia

$$\sum_{j_1, j_2, \ldots, j_h} P(j_1, j_2, \ldots, j_h) \frac{1}{h} \sum_{r=1}^{h} b_{j_r} = \frac{1}{k} \sum_{j=1}^{k} b_j.$$

In sfîrşit, adunînd relaţiile (XV.1,12) şi (XV.1,14) obţinem imediat relaţia (XV.1,7) din teorema enunţată, adică

$$D^2(Y) = \frac{k-h}{hN^2} \frac{\sum_{j=1}^{k} (\alpha_j - \alpha)}{k-1} + \frac{k}{hN^2} \sum_{j=1}^{k} N_j^2 (1-f_j) \frac{f_j^2}{n_j}.$$

Cu aceasta, teorema este complet demonstrată.

2. Teorema mai poate fi demonstrată şi direct, fără a recurge la probabilităţile condiţionate.

Fie în acest scop

$$P(j_1, j_2, \ldots, j_h)$$

probabilitatea ca, din h extracţii efectuate la întîmplare din cele k unităţi primare, să obţinem straturile

$$A_{j_1}, A_{j_2}, \ldots, A_{j_h}$$

şi

$$Y_{j_1}, Y_{j_2}, \ldots, Y_{j_h}$$

nişte variabile independente aleatorii, care corespund, respectiv, alegerii straturilor de mai sus.

Fie de asemenea

$$F_j(x) = P(Y_j \leq x) \qquad (j = 1, 2, \ldots, k)$$

funcţia de repartiţie a variabilei Y_j.

Atunci, probabilitatea ca să avem

$$x_{j_1} < Y_{j_1} < x_{j_1} + dx_{j_1},\ x_{j_2} < Y_{j_2} < x_{j_2} + dx_{j_2}, \ldots, x_{j_h} < Y_{j_h} < x_{j_h} + dx_{j_h}$$

este

$$P(j_1, j_2, \ldots, j_h) \cdot dF_{j_1}(x_{j_1}) \cdot dF_{j_2}(x_{j_2}) \ldots dF_{j_h}(x_{j_h}),$$

de unde rezultă

$$M(Y)=M\left(\frac{Y_{j_1}+Y_{j_2}+\ldots+Y_{j_h}}{h}\right)=$$

$$=\sum_{j_1,\,j_2,\ldots,\,j_h} P(j_1,\,j_2,\ldots,j_h)\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty}\frac{x_{j_1}+x_{j_2}+\ldots+x_{j_h}}{h}\,dF_{j_1}(x_{j_1})\ldots dF_{j_h}(x_{j_h})=$$

$$=\sum_{j_1,\,j_2,\ldots,\,j_h} P(j_1,\,j_2,\ldots,j_h)\,\frac{m_{j_1}+m_{j_2}+\ldots+m_{j_h}}{h}=\frac{1}{k}\sum_{j=1}^{k} m_j, \quad \text{(XV.1,15)}$$

introducînd notaţia $M(Y_j)=m_j\,(j=1,\,2,\ldots,k)$ şi ţinînd seama de valoarea mediei sondajului simplu cu o singură caracteristică.
In mod analog, obţinem

$$M(Y^2)=$$

$$=\sum_{j_1,\,j_2,\ldots,\,j_h} P(j_1,\,j_2,\ldots,j_h)\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty}\frac{(x_{j_1}+x_{j_2}+\ldots+x_{j_h})^2}{h^2}\,dF_{j_1}(x_{j_1})\ldots dF_{j_h}(x_{j_h})=$$

$$=\sum_{j_1,\,j_2,\ldots,\,j_h} P(j_1,\,j_2,\ldots,j_h)\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty}\frac{\sum_{r=1}^{h}x_{j_r}^2+2\sum_{r\neq s}x_{j_r}x_{j_s}}{h^2}\,dF_{j_1}(x_{j_1})\ldots dF_{j_h}(x_{j_h})=$$

$$=\sum_{j_1,\,j_2,\ldots,\,j_h} P(j_1,\,j_2,\ldots,j_h)\,\frac{\sum_{r=1}^{h}M_{j_r}+2\sum_{r\neq s}m_{j_r}m_{j_s}}{h^2}=$$

$$=\sum_{j_1,\,j_2,\ldots,\,j_h} P(j_1,\,j_2,\ldots,j_h)\,\frac{\sum_{r=1}^{h}\left(M_{j_r}-m_{j_r}^2\right)+\left(\sum_{r=1}^{h}m_{j_r}\right)^2}{h^2}, \quad \text{(XV.1,16)}$$

dacă notăm $M(Y_j^2)=M_j\,(j=1,\,2,\ldots,k)$.

Din relaţiile (XV.1,15) şi (XV.1,16) deducem

$$D^2(Y)=M(Y^2)-[M(Y)]^2=$$

$$=\sum_{j_1,\,j_2,\ldots,\,j_h} P(j_1,\,j_2,\ldots,j_h)\,\frac{\sum_{r=1}^{h}\left(M_{j_r}-m_{j_r}^2\right)}{h^2}+$$

$$+\sum_{j_1,\,j_2,\ldots,\,j_h} P(j_1,\,j_2,\ldots,j_h)\left(\frac{\sum_{r=1}^{h} m_{j_r}}{h}\right)^2-\left(\frac{\sum_{j=1}^{k} m_j}{k}\right)^2=$$

$$=\sum_{j_1,\,j_2,\ldots,\,j_h} P(j_1,\,j_2,\ldots,j_h)\,\frac{1}{h}\left[\frac{1}{h}\sum_{r=1}^{h}(M_{j_r}-m_{j_r}^2)\right]+$$

$$+\sum_{j_1,\,j_2,\ldots,\,j_h} P(j_1,\,j_2,\ldots,j_h)\left[\left(\frac{\sum_{r=1}^{h} m_{j_r}}{h}\right)^2-\left(\frac{\sum_{j=1}^{k} m_j}{k}\right)^2\right]=$$

$$=\frac{1}{kh}\cdot\sum_{j=1}^{k}(M_j-m_j^2)+\frac{k-h}{kh}\cdot\frac{\sum_{j=1}^{k}(m_j-m)^2}{k-1}, \qquad \text{(XV.1,17)}$$

avînd în vedere valorile mediei şi dispersiei sondajului simplu cu o singură caracteristică şi însemnînd $m=\frac{1}{k}\sum_{j=1}^{k} m_j$.

Deoarece în cazul de faţă, avem

$$m_j=M(Y_i)=\frac{k}{N}N_j\,a_j$$

şi

$$M_j-m_j^2=D^2(Y_j)=\frac{k^2}{N^2}N_j^2(1-f_j)\frac{f_j^2}{n_j},$$

relaţiile (XV.1,15) şi (XV.1,17) devin

$$M(Y)=\frac{1}{k}\sum_{j=1}^{k}\frac{k}{N}N_j\,a_j=\frac{1}{N}\sum_{j=1}^{k}N_j\,a_j=a,$$

$$D^2(Y)=$$

$$=\frac{k-h}{kh}\cdot\frac{\sum_{j=1}^{k}\left(\frac{k}{N}N_j a_j-\frac{1}{k}\sum_{j=1}^{k}\frac{k}{N}N_j a_j\right)^2}{k-1}+\frac{1}{kh}\sum_{i=1}^{k}\frac{k^2}{N^2}N_i^2(1-f_j)\frac{f_j^2}{n_j}=$$

$$=\frac{k(k-h)}{hN^2}\cdot\frac{\sum_{j=1}^{k}(\alpha_j-\alpha)^2}{k-1}+\frac{k}{hN^2}\sum_{j=1}^{k}N_j^2(1-f_j)\frac{f_j^2}{n_j}.$$

Am regăsit, astfel, relaţiile (XV.1,6) şi (XV.1,7) din teoremă.

Observaţii. După cum se constată, dispersia variabilei aleatorii Y este suma a doi termeni

$$D_1=\frac{k(k-h)}{hN^2}\cdot\frac{\sum_{j=1}^{k}(\alpha_j-\alpha)^2}{k-1}$$

şi

$$D_2=\frac{k}{hN^2}\cdot\sum_{j=1}^{k}N_j^2(1-f_j)\frac{f_j^2}{n_j}.$$

Primul termen D_1 este dispersia corespunzătoare primei faze a sondajului şi se referă la suma caracteristicilor unităţilor secundare $\alpha_j=\sum_{i=1}^{N_j}a_{ij}$. Deci, putem considera acest termen ca o dispersie între straturi. El se anulează cînd $h=k$, adică atunci cînd sondajul încetează de a mai fi cu două faze şi se reduce astfel la un sondaj stratificat simplu. El se mai anulează de asemenea şi cînd

$$\alpha_1=\alpha_2=\ldots=\alpha_k,$$

cu alte cuvinte atunci cînd suma caracteristicilor unităţilor secundare este aceeaşi.

Al doilea termen D_2 reprezintă dispersia din interiorul straturilor, care corespund fazei a doua a sondajului. El se anulează cînd avem $n_j=N_j$ $(j=1, 2,\ldots, k)$, ceea ce înseamnă atunci cînd scoatem toate unităţile secundare din fiecare strat.

Rezultă de aici că, dacă numerele n_j sînt suficient de mari, termenul $D_2\to 0$, în timp ce termenul D_1 rămîne constant, deoarece el nu depinde de n_j.

Prin urmare, variabila aleatorie

$$Y=\frac{Y_{j_1}+Y_{j_2}+\ldots+Y_{j_h}}{h}$$

nu este o funcţie de estimaţie pentru media caracteristicilor colectivităţii totale a, oricît de mare ar fi mărimea $n=\sum_{i=1}^{k}n_j$ a sondajului.

Scanned with OKEN Scanner

Pentru ca din punct de vedere practic să ajungem la rezultate acceptabile printr-un sondaj cu două faze, este necesar ca dispersia $D^2(Y)$ să fie foarte apropiată de zero pentru numerele n_j $(j=1, 2,..., k)$ suficient de mari. Aceasta, însă, după cele arătate mai sus, nu se poate obține decît numai dacă D_1 este, la rîndul său, foarte apropiat de zero. Insă, pentru ca D_1 să fie egal cu zero, trebuie ca $h=k$ sau $\alpha_j=\alpha$ $(j=1, 2,..., k)$. Primul caz nu interesează, deoarece avantajul sondajului cu două faze constă tocmai în faptul de a alege la întîmplare un număr mai mic de unități primare, decît cele existente.

Rămîne deci de rezolvat numai cazul $\alpha_j=\alpha$ $(j=1, 2,..., k)$, de unde rezultă că spre deosebire de sondajul stratificat simplu, sondajul cu două faze nu este aplicabil decît după o prealabilă prelucrare a straturilor.

Pentru a pune în evidență erorile ce se pot ivi la aplicarea fără un control prealabil a sondajului cu două faze, dăm următorul exemplu, redus la o formă schematică simplă.

Presupunem, astfel, că avem

$$k=4, \quad h=2,$$

adică un sondaj cu patru straturi din care am ales la întîmplare cîte două.

Admitem că pentru fiecare strat avem

$N_1=10\,000,$	$\alpha_1=100\,000$
$N_2=12\,000,$	$\alpha_2=\ 25\,000$
$N_3=10\,000,$	$\alpha_3=200\,000$
$N_4=18\,000,$	$\alpha_4=\ 75\,000$
$N\ =50\,000$	$\alpha\ =400\,000.$

Valorile α_i diferă sensibil între ele.

Media caracteristicilor colectivității totale este

$$a=\frac{\alpha}{N}=8.$$

Să mai admitem că, după ce am ales cele două unități primare, lucrăm în cele mai exacte condiții, adică din fiecare strat scoatem toate unitățile secundare. In aceste condiții, aplicînd formulele stabilite, rezultă pentru media a următoarele estimații

Faza a doua a sondajului s-a făcut din straturile	*Estimația mediei* a
A_1, A_2	5
A_1, A_3	12
A_1, A_4	7
A_2, A_3	9
A_2, A_4	4
A_3, A_4	11.

Valoarea exactă a mediei a fiind 8, se vede de aici că aplicarea sondajului cu două faze în cele mai optime condiții ne conduce la rezultate inacceptabile din punct de vedere practic.

Să presupunem că, păstrînd aceleași unități primare și secundare, avem

$N_1 = 10\,000,$	$\alpha_1 = 100\,000$
$N_2 = 12\,000,$	$\alpha_2 = 90\,000$
$N_3 = 10\,000,$	$\alpha_3 = 110\,000$
$N_4 = 18\,000,$	$\alpha_4 = 100\,000$
$N = 50\,000$	$\alpha = 400\,000.$

In acest caz, valorile α_i nu diferă prea mult între ele și deci ne așteptăm la estimații mult mai apropiate de realitate.

Media caracteristicilor colectivității totale este aceeași

$$a = \frac{\alpha}{N} = 8.$$

Lucrînd în aceleași condiții, adică alegînd la întîmplare două straturi din cele patru straturi și extrăgînd din fiecare strat toate unitățile secundare, obținem pentru media a estimațiile de mai jos.

Faza a doua a sondajului s-a făcut din straturile	*Estimația mediei* a
A_1, A_2	7,6
A_1, A_3	8,8
A_1, A_4	8,0
A_2, A_3	8,0
A_2, A_4	7,6
A_3, A_4	8,8

Aceste rezultate sînt mult mai apropiate de valoarea exactă opt. Este evident că, dacă am lua toate valorile

$$\alpha_i = 100\,000 \qquad (i = 1, 2, 3, 4),$$

estimația medie a ar fi chiar egală cu opt.

Această observație are o valoare practică, deoarece ne arată că, pentru aplicarea unui sondaj cu două faze, este necesar să ne alegem astfel straturile, încît suma caracteristicilor unităților din fiecare strat să fie aproximativ aceeași.

O altă observație avem de făcut atunci cînd în practică se aplică sondajul stratificat reprezentativ. In acest caz avem, după cum se știe, relațiile

$$\frac{n_j}{N_j} = \frac{n}{N} \qquad (j = 1, 2, \ldots, k).$$

Atunci, formulele (XV.1,6) și (XV,1,7) se pot pune sub forma mai simplă

$$M(Y)_k = a$$

și

$$D^2(Y) = \frac{k(k-h)}{hN^2} \cdot \frac{\sum_{j=1}^{k} (\alpha_j - \alpha)^2}{k-1} + \frac{k(N-n)}{hnN^2} \sum_{j=1}^{k} N_j f_j^2,$$

unde

$$n = n_1 + n_2 + \ldots + n_k,$$
$$N = N_1 + N_2 + \ldots + N_k.$$

§ 2. Sondajul cu două faze cu mai multe caracteristici

Să examinăm problema mai generală a sondajului cu două faze pentru cazul cînd unităţile fiecărui strat A_j $(j=1, 2, \ldots, k)$ prezintă fiecare cîte două caracteristici.

Cu alte cuvinte, unităţile secundare sînt respectiv, vectorii aleatori

$$l_{ij}(a_{ij}, b_{ij}) \qquad (i=1, 2, \ldots, N_j) \qquad (j=1, 2, \ldots, k),$$

iar unităţile corespunzătoare fazei a doua a sondajului sînt, respectiv, vectorii aleatori

$$x_{ij}(u_{ij}, v_{ij}) \qquad (i=1, 2, \ldots, n_j) \qquad (j=1, 2, \ldots, k).$$

Fie atunci

$$Y_j = \frac{kN_j}{N} \cdot \frac{\sum_{i=1}^{n_j} x_{ij}}{n_j},$$

vectorul aleator care are componentele u_j şi v_j şi

$$Y = \frac{Y_{j_1} + Y_{j_2} + \ldots + Y_{j_h}}{h}$$

vectorul aleator cu componentele

$$U = \frac{u_{j_1} + u_{j_2} + \ldots + u_{j_h}}{h}$$

şi

$$V = \frac{v_{j_1} + v_{j_2} + \ldots + v_{j_h}}{h}.$$

Toate notaţiile introduse la studiul sondajului stratificat simplu cu două caracteristici rămîn valabile şi pentru sondajul cu două faze, avînd două caracteristici.

Teoremă. *Intr-un sondaj cu două faze, avînd două caracteristici, există relaţiile*

$$M(U)=a, \quad M(V)=b,$$

$$D^2(U)=\frac{k(k-h)}{hN^2}\cdot\frac{\sum_{j=1}^{k}(\alpha_j-\alpha)^2}{k-1}+\frac{k}{hN^2}\sum_{j=1}^{k}N_j^2(1-f_j)\frac{f_{a_ja_j}^2}{n_j},$$

$$D^2(V)=\frac{k(k-h)}{hN^2}\cdot\frac{\sum_{j=1}^{k}(\beta_j-\beta)^2}{k-1}+\frac{k}{hN^2}\sum_{j=1}^{k}N_j^2(1-f_i)\frac{f_{b_jb_j}^2}{n_j},$$

$$C(UV)=\frac{k(k-h)}{hN^2}\cdot\frac{\sum_{j=1}^{k}(\alpha_j-\alpha)(\beta_j-\beta)}{k-1}+\frac{k}{hN^2}\sum_{j=1}^{k}N_j^2(1-f_j)\frac{f_{a_jb_j}^2}{n_j}, \tag{XV.2,1}$$

unde am notat prin $C(UV)=M(U-a)(V-b)$ *covarianţa variabilelor* U *şi* V, *iar*

$$\alpha_j=\sum_{i=1}^{N_j}a_{ij}, \quad \beta_j=\sum_{i=1}^{N_j}b_{ij}, \quad \alpha=\frac{1}{k}\sum_{j=1}^{k}a_j \quad \text{şi} \quad \beta=\frac{1}{k}\sum_{j=1}^{k}b_j.$$

Primele patru formule din (XV.2,1) se stabilesc întocmai ca la sondajul cu două faze cu o singură caracteristică. Ultima formulă se poate demonstra folosind lema 2, adică bazîndu-ne pe considerarea mediilor condiţionate.

In cele ce urmează, vom da însă o demonstraţie directă.

Să notăm cu

$$P(j_1, j_2, \ldots, j_h)$$

probabilitatea ca în prima fază a sondajului să obţinem straturile

$$A_{j_1}, A_{j_2}, \ldots, A_{j_h}$$

şi cu

$$F_j(x, y)=P(U_j \leqq x, \; V_j \leqq y) \qquad (j=1, 2, \ldots, k),$$

funcţia de repartiţie a vectorului $Y_j(U_j, V_j)$.

Atunci, probabilitatea de a avea

$$x_{j_1} < U_{j_1} < x_{j_1} + dx_{j_1},\ x_{j_2} < U_{j_2} < x_{j_2} + dx_{j_2}, \ldots, x_{j_h} < U_{j_h} < x_{j_h} + dx_{j_h}$$

şi

$$y_{j_1} < V_{j_1} < y_{j_1} + dy_{j_1},\ y_{j_2} < V_{j_2} < y_{j_2} + dy_{j_2}, \ldots, y_{j_h} < V_{j_h} < y_{j_h} + dy_{j_h}$$

este egală cu

$$P(j_1, j_2, \ldots, j_h)\, dF_{j_1}(x_{j_1}, y_{j_1})\, dF_{j_2}(x_{j_2}, y_{j_2}) \ldots dF_{j_h}(x_{j_h}, y_{j_h}),$$

de unde rezultă

$$M(UV) = \sum_{j_1, j_2, \ldots, j_h} P(j_1, j_2, \ldots, j_h) \iint \ldots \int \frac{x_{j_1} + x_{j_2} + \ldots + x_{j_h}}{h} \times$$

$$\times \frac{y_{j_1} + y_{j_2} + \ldots + y_{j_h}}{h}\, dF_{j_1}(x_{j_1}, y_{j_1}) \cdot dF_{j_2}(x_{j_2}, y_{j_2}) \ldots dF_{j_h}(x_{j_h}, y_{j_h}), \tag{XV.2,2}$$

fiecare dintre integrale extinzîndu-se în tot planul.

Să mai introducem notaţia

$$m_j^{(1)} = \int x_j dF_j(x_j, y_j),$$

$$m_j^{(2)} = \int y_j dF_j(x_j, y_j)$$

şi

$$M_{ij} = \int x_j y_j dF_j(x_j, y_j),$$

unde fiecare dintre aceste integrale se extinde, la fel, în tot planul. Cu aceste notaţii, relaţia (XV.2,2) devine

$$M(UV) = \sum_{j_1, j_2, \ldots, j_h} P(j_1, j_2, \ldots, j_h) \frac{\sum_{r=1}^{h} M_{j_r j_r} + \sum_{r \neq s} m_{j_r}^{(1)} m_{j_s}^{(2)}}{h^2} =$$

$$= \sum_{j_1, j_2 \ldots, j_h} P(j_1, j_2, \ldots, j_h) \frac{\sum_{r=1}^{h} (M_{j_r j_r} - m_{j_r}^{(1)} m_{j_r}^{(2)}) + \sum_{r=1}^{h} m_{j_r} \sum_{s=1}^{h} m_{j_s}}{h^2}. \tag{XV.2,3}$$

Pe de altă parte, avem

$$M(U)\cdot M(V)=\frac{\sum_{j=1}^{k} m_j^{(1)}}{k}\cdot\frac{\sum_{j=1}^{k} m_j^{(2)}}{k}\cdot \qquad \text{(XV.2,4)}$$

Din relaţiile (XV.2,3) şi (XV.2,4), deducem

$$C(UV)=M(UV)-M(U)\,M(V)=$$

$$=\sum_{j_1, j_2, \ldots, j_h} P(j_1, j_2, \ldots, j_h)\,\frac{\sum_{r=1}^{h}(M_{j_r j_r}-m_{j_r}^{(1)} m_{j_r}^{(2)})}{h^2}+$$

$$+\sum_{j_1, j_2, \ldots, j_h} P(j_1, j_2, \ldots, j_h)\left[\frac{\sum_{r=1}^{h} m_{j_r}}{h}\cdot\frac{\sum_{s=1}^{h} m_{j_s}}{h}-\frac{\sum_{j=1}^{k} m_j^{(1)}}{k}\cdot\frac{\sum_{j=1}^{k} m_j^{(2)}}{k}\right]=$$

$$=\frac{\sum_{j=1}^{k}(M_{jj}-m_j^{(1)} m_j^{(2)})}{hk}+\frac{k-h}{hk}\,\frac{\sum_{j=1}^{k}(m_j^{(1)}-m^{(1)})(m_j^{(2)}-m^{(2)})}{k-1}, \qquad \text{(XV.2,5)}$$

avînd în vedere formulele stabilite la studiul sondajului simplu cu două caracteristici şi însemnînd

$$m^{(1)}=\frac{1}{k}\sum_{j=1}^{k} m_j^{(1)},\quad m^{(2)}=\frac{1}{k}\sum_{j=1}^{k} m_j^{(2)}.$$

Observînd că, în cazul de faţă, putem scrie

$$M_{jj}-m_j^{(1)} m_j^{(2)}=\frac{k^2 N_j^2}{N^2}(1-f_j)\,\frac{s_{a_j b_j}^2}{n_j}$$

şi

$$(m_j^{(1)}-m^{(1)})(m_j^{(2)}-m^{(2)})=\frac{k^2(\alpha_j-\alpha)(\beta_j-\beta)}{N^2},$$

rezultă, prin înlocuire în (XV.2,5), ultima formulă din relaţiile (XV.2,1) din teorema enunţată, adică

$$C(UV)=\frac{k(k-h)}{hN^2}\cdot\frac{\sum_{j=1}^{k}(\alpha_j-\alpha)(\beta_j-\beta)}{k-1}+\frac{k}{hN^2}\cdot\sum_{j=1}^{k} N_j^2(1-f_j)\frac{s_{a_j b_j}^2}{n_j}\cdot$$

Scanned with OKEN Scanner

Toate problemele care se pun în legătură cu sondajul cu o singură fază se pun şi pentru sondajul cu două faze.

De exemplu, ne putem propune estimarea raportului $\frac{a}{b}$ dintre mediile caracteristicilor colectivităţii totale, precum şi problema determinării unui sondaj optim, ţinînd seama de eficienţa sondajului sau de cheltuielile totale.

Aceste probleme nu le vom trata, ele putînd fi tratate prin metode analoge cu cele folosite la studiul sondajului cu o singură fază.

Scanned with OKEN Scanner

PARTEA A TREIA

Capitolul XVI

Teoria estimației pentru familiile de distribuții cu un parametru

§ 1. Problema estimației parametrilor

In unele aplicații ale statisticii matematice, în tehnică sau științele experimentale, există motive teoretice pentru a putea admite că distribuția fenomenului care ne interesează este dată de o funcție cunoscută, în care intră anumiți parametri cu valori necunoscute. De exemplu, aplicînd teorema lui Liapunov unui fenomen care îndeplinește toate condițiile acestei teoreme, ajungem la concluzia că distribuția fenomenului este normală. Pentru aplicații practice, nu ne putem însă mulțumi numai cu atît, ci trebuie să cunoaștem distribuția exactă. Cu alte cuvinte, trebuie să determinăm valorile numerice ale celor doi parametri m și σ care intră în alcătuirea distribuției normale

$$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}}.$$

Un exemplu din statistica practică de estimație a parametrilor este dat de construcția unei tabele de mortalitate. Se știe că o tabelă de mortalitate dă valorile probabilităților p_x, ca un individ de vîrstă x să fie în viață la împlinirea vîrstei de $x+1$ ani. Presupunem că această probabilitate este aceeași pentru toți indivizii de vîrstă x. Rezultă că sîntem în prezența unei distribuții binomiale cu probabilitatea p_x a cărei valoare nu o cunoaștem și ne propunem să o calculăm pe baza datelor experimentale pe care le avem la dispoziție.

Vom spune despre distribuția exprimată printr-o funcție dată, în care intră anumiți parametri necunoscuți, că reprezintă o lege de distribuție *specificată*. In cazul cînd cunoaștem valorile numerice

luate de parametru, deci cînd distribuţia este cunoscută în întregime, avem o lege de distribuţie *complet specificată*.

Determinarea valorilor numerice ale parametrilor unei distribuţii specificate se face folosind un număr, să zicem n, de experienţe independente, care ne conduc la valorile $x_1, x_2, \ldots, x_n$ pentru variabila statistică pe care o studiem. Valorile x_i reprezintă după o denumire des folosită în statistica matematică, o valoare informativă pentru determinarea valorilor adevărate ale parametrilor legii de distribuţie.

Independenţa celor n experienţe $x_1, x_2, \ldots, x_n$ rezultă din natura fenomenului statistic considerat. Sînt cazuri cînd aceste variabile sînt dependente. Problema estimării parametrilor se pune şi în acest caz, după cum vom vedea mai tîrziu.

Există mai multe metode pentru estimaţia parametrilor unei distribuţii specificate. Ne vom ocupa de distribuţiile specificate care depind de un singur parametru.

Fie $f(x, \alpha)$ o familie de funcţii de frecvenţă a unor distribuţii specificate continue, α fiind un parametru cu valori reale. Expunem metodele cele mai importante de estimaţie pentru astfel de familii de distribuţii statistice continue. Aceste metode sînt aplicabile fără deosebiri esenţiale şi pentru familiile de distribuţii statistice discontinuie, pentru care vom cita numai rezultatele, lăsînd demonstraţiile în seama cititorului.

Admitem continuitatea funcţiilor $f(x, \alpha)$ în R, existenţa derivatelor acestor funcţii în raport cu x şi α, pînă la ordinele necesare în calculele care vor urma, precum şi existenţa diferitelor momente necesare în aceste calcule. Aceste condiţii sînt presupuse verificate în toate teoremele care urmează, fără ca să mai indicăm aceasta de fiecare dată.

§ 2. Aplicarea teoremei lui Bayes

Estimaţia parametrilor unei legi de distribuţie specificată este o problemă în care folosirea teoremei lui Bayes se impune în mod natural.

Intr-adevăr, determinarea valorii numerice a parametrilor, care intră în alcătuirea legii de distribuţie, în baza unui număr oarecare de experienţe, revine la găsirea cauzei care a condus la rezultatul acestor experienţe. Este evident că problema nu admite o soluţie determinată.

Nu este vorba de a stabili valoarea parametrului în funcție de datele selecției $x_1, x_2, \ldots, x_n$, ci numai probabilitatea ca acest parametru să aibă o anumită valoare. Analiza formulei lui Bayes ne arată că trebuie să fim foarte prudenți la aplicarea acestei formule, deoarece trebuie să admitem că valoarea parametrului este aleatorie, pe cînd în realitate ea are o valoare bine determinată, pe care însă noi nu o cunoaștem. Admiterea acestei ipoteze, necorespunzătoare realității, a făcut ca formula lui Bayes să intre în oarecare măsură în umbră pentru a lăsa locul unor metode directe de estimație. Vom începe totuși acest capitol cu expunerea metodei bazate pe aplicarea formulei lui Bayes, deoarece în ultima vreme a căpătat din ce în ce mai multe întrebuințări.

Fie $f(x, \alpha)$ densitatea de repartiție a legii specificate, în care intră un singur parametru α cu valori reale, $g(\alpha)$ densitatea de repartiție a lui α și $g((\alpha \mid \vec{x}) = g(\alpha))$ densitatea de repartiție condiționată a lui α, în ipoteza că am obține în cele n experiențe independente grupul de valori $x_1, x_2, \ldots, x_n$.

Potrivit formulei lui Bayes,

$$g(\alpha|\vec{x}) = \frac{f(x_1, \alpha) f(x_2, \alpha) \ldots f(x_n, \alpha)}{\int_{-\infty}^{+\infty} f(x_1, \alpha) f(x_2, \alpha) \ldots f(x_n, \alpha)\, d\vec{x}}.$$

Parametrul considerat este o variabilă aleatorie, a cărei lege de probabilitate depinde de observațiile făcute $x_1, x_2, \ldots, x_n$. După cîte știm, legea de probabilitate nu ne dă valoarea exactă pe care o ia variabila respectivă; ea ne poate da însă indicații asupra repartiției valorilor pe care le ia această variabilă.

De exemplu, probabilitatea inegalității $a \leqq \alpha \leqq b$, unde a și b sînt două numere reale date, este

$$P(a \leqq \alpha \leqq b) = \int_a^b g(\alpha \mid \vec{x})\, d\alpha.$$

Dacă $P(a \leqq \alpha \leqq b)$ este foarte aproape de unu putem afirma că, din punct de vedere practic, parametrul se găsește cuprins în intervalul (a, b), valabilitatea acestei afirmații fiind în funcție de mărimea probabilității $P(a \leqq \alpha \leqq b)$.

Se vede că, în mod obișnuit, prin acest procedeu nu obținem o estimație pentru α, ci numai un interval în care, cu certitudine prac-

tică, parametrul α variază. Cu cît acest interval este mai mic, cu atît informaţia noastră asupra adevăratei valori a lui α este mai precisă.

Metoda nu este deci aplicabilă decît în cazurile particulare cînd există o concentrare a masei de probabilitate într-o regiune restrînsă.

Vom da cîteva exemple.

1. *Urna lui Bernoulli.* In lucrările clasice de calculul probabilităţilor găsim tratată, cu ajutorul formulei lui Bayes, determinarea probabilităţii necunoscute p din urna lui Bernoulli, din care făcînd n extracţii consecutive am obţinut frecvenţa $\frac{\nu}{n}$. In cazul de faţă parametrul α, care intră în alcătuirea legii de distribuţie specificate, este p, adică valoarea probabilităţii constante de a se verifica evenimentul în fiecare probă.

Valorile observate $x_i\,(i=1, 2, \ldots, n)$ pot lua valorile unu sau zero, respectiv, după cum evenimentul s-a verificat sau nu, în fiecare probă. Ştim însă că

$$x_1+x_2+\ldots+x_n=\nu.$$

Funcţia $f(x_i, \alpha)$ ia valoarea p sau $1-p$, după cum $x_i=1$ sau $x_i=0$. Sîntem deci în cazul unei distribuţii discontinue. Metoda rămîne evident aceeaşi.

Aplicînd formula care ne dă probabilitatea condiţionată a lui $\alpha=p$ faţă de $x_i\,(i=1, 2, \ldots, n)$, obţinem

$$g(p\,|\,\vec{x})=\frac{p^{\nu}q^{n-\nu}g(p)}{\int_0^1 p^{\nu}q^{n-\nu}g(p)\,dp}.$$

Să observăm că, în cel de al doilea membru al egalităţii de mai sus, nu intră decît $\sum_{i=1}^{n} x_i=\nu$, ceea ce înseamnă că $\sum_{i=1}^{n} x_i$ este o statistică exhaustivă pentru estimaţia lui p.

Probabilitatea $g(p\,|\,\vec{x})$ depinde de probabilitatea a priori $g(p)$.

Să facem deocamdată ipoteza cea mai simplă că $g(p)=1$. Atunci

$$g(p\,|\,\vec{x})=\frac{p^{\nu}q^{n-\nu}}{\int_0^1 p^{\nu}q^{n-\nu}dp}.$$

Insă

$$\int_0^1 p^\nu q^{n-\nu} dp = \int_0^1 p^\nu (1-p)^{n-\nu} dp = B(\nu+1,\ n-\nu+1) = \frac{(n+1)!}{\nu!\,(n-\nu)!}.$$

Rezultă că

$$g(p \mid \vec{x}) = \frac{(n+1)!}{\nu!\,(n-\nu)!} p^\nu (1-p)^{n-\nu}.$$

Să calculăm valoarea medie şi dispersia variabilei p în raport cu această lege de probabilitate. Avem

$$M(p) = \frac{(n+1)!}{\nu!\,(n-\nu)!} \int_0^1 p^{\nu+1} (1-p)^{n-\nu} dp = \frac{(n+1)!}{\nu!\,(n-\nu)!} B(\nu+2,\ n-\nu+1) =$$

$$= \frac{(n+1)!}{\nu!\,(n-\nu)!} \cdot \frac{(\nu+1)!\,(n-\nu)!}{(n+2)!} = \frac{\nu+1}{n+2}.$$

In mod analog obţinem

$$D^2(p) = \frac{(\nu+1)(n-\nu+1)}{(n+2)^2 (n+3)}.$$

Dispersia $D^2(p)$ ia cea mai mare valoare pentru $\nu = \frac{n}{2}$. Avem deci

$$D^2(p) \leq \frac{1}{4(n+3)},$$

pentru orice ν cuprins între 0 şi n.
După inegalitatea lui Cebîşev,

$$P\left(\left|p - \frac{\nu+1}{n+2}\right| < \varepsilon\right) > 1 - \frac{1}{4(n+3)\varepsilon^2}.$$

Rezultă că avem o concentrare a masei de probabilitate în punctul $\frac{\nu+1}{n+2}$, care poate fi considerat, astfel, ca o estimaţie a lui p, pentru valorile mari date lui n.
Ipoteza $f(p)=1$ este prea restrictivă. Se poate arăta că, dacă presupunem $f(p)$ mărginită superior de un număr L şi inferior de un număr $l>0$, rezultatul obţinut mai sus se păstrează.

Scanned with OKEN Scanner

Legea de probabilitate condiţionată a lui p este în acest caz

$$g(p \mid \nu) = \frac{p^{\nu}(1-p)^{n-\nu} f(p)}{\int_0^1 p^{\nu}(1-p)^{n-\nu} f(p)\, dp}.$$

Prin urmare

$$P\left(\left|p - \frac{\nu+1}{n+2}\right| < \varepsilon\right) = \frac{\int_{\frac{\nu+1}{n+2}-\varepsilon}^{\frac{\nu+1}{n+2}+\varepsilon} p^{\nu}(1-p)^{n-\nu} f(p)\, dp}{\int_0^1 p^{\nu}(1-p)^{n-\nu} f(p)\, dp} =$$

$$= 1 + \frac{\int_0^{\frac{\nu+1}{n+2}-\varepsilon} p^{\nu}(1-p)^{n-\nu} f(p)\, dp + \int_{\frac{\nu+1}{n+2}+\varepsilon}^{n} p^{\nu}(1-p)^{n-\nu} f(p)\, dp}{\int_{\frac{\nu+1}{n+2}-\varepsilon}^{\frac{\nu+1}{n+2}+\varepsilon} p^{\nu}(1-p)^{n-\nu} f(p)\, dp}.$$

Insă

$$\frac{\int_0^{\frac{\nu+1}{n+2}-\varepsilon} p^{\nu}(1-p)^{n-\nu} f(p)\, dp + \int_{\frac{\nu+1}{n+2}+\varepsilon}^{n} p^{\nu}(1-p)^{n-\nu} f(p)\, dp}{\int_{\frac{\nu+1}{n+2}-\varepsilon}^{\frac{\nu+1}{n+2}+\varepsilon} p^{\nu}(1-p)^{n-\nu} f(p)\, dp} <$$

$$< \frac{L}{l} \cdot \frac{P\left(\left|p - \frac{\nu+1}{n+2}\right| > \varepsilon\right)}{P\left(\left|p - \frac{\nu+1}{n+2}\right| < \varepsilon\right)},$$

ultimele probabilităţi referindu-se la cazul $f(p) = 1$, pentru care am calculat şi valoarea medie şi dispersia $D^2(p)$.

Aplicînd inegalitatea lui Cebîşev

$$\frac{L}{l}\cdot\frac{P\left(\left|p-\frac{\nu+1}{n+2}\right|>\varepsilon\right)}{P\left(\left|p-\frac{\nu+1}{n+2}\right|<\varepsilon\right)}<\frac{L}{l}\cdot\frac{\frac{D^2(p)}{\varepsilon^2}}{1-\frac{D^2(p)}{\varepsilon^2}}<\frac{L}{l}\cdot\frac{1}{4(n+3)\varepsilon^2-1},$$

rezultă

$$P\left(\left|p-\frac{\nu+1}{n+2}\right|<\varepsilon\right)\geqq\frac{1}{1+\frac{L}{l}\cdot\frac{1}{4(n+3)\varepsilon^2-1}}.$$

Oricît de mic ar fi $\varepsilon>1$, înseamnă că pentru valori mari ale lui n avem o concentrare a masei de probabilitate în jurul punctului $\frac{\nu+1}{n+2}$ care deci poate fi considerat ca o estimație a lui p în ipoteza destul de largă pe care am făcut-o asupra probabilității a priori a lui p. Asimptotic $\frac{\nu+1}{n+2}$ nu diferă de frecvența $\frac{\nu}{n}$.

2. *Legea normală.* Fie

$$f(x;\alpha)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\alpha)^2}{2\sigma^2}},$$

unde σ este dat, iar α este un parametru, a cărei valoare avem s-o determinăm în funcție de n observații independente $x_1, x_2, \ldots, x_n$. Aplicînd formulele găsite, rezultă pentru probabilitatea lui α condiționată de x_i

$$g(\alpha\,|\,\vec{x})=\frac{e^{-\frac{\sum_{i=1}^{n}(x_i-\alpha)^2}{2\sigma^2}}}{\int_{-\infty}^{+\infty}e^{-\frac{\sum_{i=1}^{n}(x_i-\alpha)^2}{2\sigma^2}}g(\alpha)\,d\alpha},$$

Insă

$$\frac{\sum_{i=1}^{n}(x_i-\alpha)^2}{2\sigma^2}=\sum_{i=1}^{n}\frac{(x_i-\bar{x})^2}{2\sigma^2}=\frac{n(\bar{x}-\alpha)^2}{2\sigma^2},$$

unde am notat cu $\bar{x}$, ca de obicei, media aritmetică a valorilor x_i.

Rezultă că

$$g(\alpha \mid \vec{x}) = \frac{e^{-\frac{n(\bar{x}-\alpha)^2}{2\sigma^2}}}{\int_{-\infty}^{+\infty} e^{-\frac{n(\bar{x}-\alpha)^2}{2\sigma^2}} g(\alpha)\, d\alpha}.$$

Se vede că pentru determinarea probabilității $g(\alpha \mid \vec{x})$, care folosește la estimația lui α, grupul valorilor $x_1, x_2, \ldots, x_n$ nu intervine decît prin funcția $\bar{x}$, care constituie deci o statistică exhaustivă pentru estimația lui α în raport cu grupul de valori $x_1, x_2, \ldots, x_n$.

Se poate demonstra ușor că, dacă funcția $g(\alpha)$ admite o derivată mărginită și ia o valoare diferită de zero în punctul $\bar{x}$, probabilitatea $g(\alpha \mid \vec{x}) = g(\alpha \mid x_1, x_2, \ldots, x_n)$ tinde către distribuția normală $\left(x, \frac{\sigma}{\sqrt{n}}\right)$, cînd $n \to \infty$. Prin urmare, pentru valorile mari ale lui n avem o concentrare a masei de repartiție $g(\alpha \mid \vec{x})$ în jurul valorii $\bar{x}$.

Într-adevăr, funcția caracteristică a distribuției lui α, condiționată de valorile x_i, este

$$f_n(t) = M\left(g(\alpha \mid \vec{x})\, e^{i\alpha t}\right) = \frac{\int_{-\infty}^{+\infty} e^{-\frac{n(\bar{x}-\alpha)^2}{2\sigma^2}} g(\alpha)\, e^{i\alpha t}\, d\alpha}{\int_{-\infty}^{+\infty} e^{-\frac{n(\bar{x}-\alpha)^2}{2\sigma^2}} g(\alpha)\, d\alpha}.$$

Să facem substituția

$$\alpha = \bar{x} + \frac{z\sigma}{\sqrt{n}}.$$

Rezultă atunci că

$$f_n(t) = \frac{\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{z^2}{2}} g\left(\bar{x} + \frac{z\sigma}{\sqrt{n}}\right) e^{\left(\bar{x} + \frac{z\sigma}{\sqrt{n}}\right) ti}\, dz}{\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{z^2}{2}} g\left(\bar{x} + \frac{z\sigma}{\sqrt{n}}\right) dz}.$$

Insă

$$g\left(\bar{x}+\frac{z\sigma}{\sqrt{n}}\right)=g(\bar{x})+\frac{z\sigma}{\sqrt{n}}g'(\tau),$$

unde

$$g(\bar{x})\neq 0,$$

$$\tau=\bar{x}+\theta\frac{z\sigma}{\sqrt{n}}\quad(|\theta|<1),$$

$$|g'(\tau)|<L<\infty.$$

Prin urmare

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}e^{-\frac{z^2}{2}}g(\bar{x})\,e^{\left(\bar{x}+\frac{z\sigma}{\sqrt{n}}\right)ti}\,dz=\frac{1}{\sqrt{2\pi}}g(\bar{x})\,e^{\bar{x}ti}\int_{-\infty}^{+\infty}e^{-z^2}e^{\frac{z\sigma}{\sqrt{n}}ti}\,dz=$$

$$=g(\bar{x})\,e^{\bar{x}ti-\frac{1}{2}\frac{\sigma^2}{n}t^2},$$

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}e^{-\frac{z^2}{2}}g(\bar{x})\,dz=g(\bar{x}),$$

$$\left|\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}e^{-\frac{z^2}{2}}\frac{z\sigma}{\sqrt{n}}g'(\tau)\,dz\right|\leq\frac{2}{\sqrt{2\pi}}\cdot\frac{\sigma}{\sqrt{n}}L\int_{0}^{\infty}ze^{-\frac{z^2}{2}}dz=\frac{\sigma\sqrt{2}}{\sqrt{\pi}\sqrt{n}}L,$$

$$\left|\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}e^{-\frac{z^2}{2}}\frac{z\sigma}{\sqrt{n}}g'(\tau)\,e^{\left(\bar{x}+\frac{z\sigma}{\sqrt{n}}\right)ti}\,dz\right|\leq\frac{\sigma\sqrt{2}}{\sqrt{\pi}\sqrt{n}}L.$$

Obținem

$$f_n(t)=e^{\bar{x}ti-\frac{1}{2}\frac{\sigma^2}{n^2}t^2}+O\left(\frac{1}{\sqrt{n}}\right).$$

Cum $e^{\bar{x}ti-\frac{1}{2}\frac{\sigma^2}{n}t^2}$ este funcția caracteristică a distribuției normale $\left(\bar{x},\frac{\sigma}{\sqrt{n}}\right)$, teorema este demonstrată. Putem deci să considerăm pe $\bar{x}$ ca o estimație a lui α.

§ 3. Metode directe de estimație

Să considerăm din nou familia de legi de distribuție $f(x, \alpha)$, unde α este un parametru. Admitem că am făcut n experiențe, care ne-au condus la valorile $x_1, x_2, \ldots, x_n$. Fie $\bar{\alpha} = \alpha(\vec{x}) = \alpha(x_1, x_2, \ldots, x_n)$ o funcție de cele n observații $x_1, x_2, \ldots, x_n$.

Evaluăm valoarea necunoscută a parametrului α în mod direct, folosind o clasă de funcții $\alpha(x)$ care se bucură de unele proprietăți, pe care le vom arăta mai jos. Aceste funcții poartă numele de *funcții de estimație.*

Este evident că, pentru ca $\alpha(\vec{x}) = \alpha(x_1, x_2, \ldots, x_n)$ să fie o funcție de selecție, este necesar ca pentru valorile mari ale lui n, $\alpha(\vec{x})$ să conveargă în probabilitate către α.

Valoarea luată de $\alpha(x_1, x_2, \ldots, x_n)$ pentru valori bine determinate ale variabilelor $x_1, x_2, \ldots, x_n$ poartă numele de *estimație* a lui α. Deci trebuie să facem o diferență între funcția de estimație, care este teoretică și estimația, care este de natură empirică.

Dacă $\bar{\alpha}$ tinde în probabilitatea către α cînd $n \to \infty$, vom spune că $\bar{\alpha}$ reprezintă o *funcție de estimație corectă* a lui α. Convergența în probabilitate a lui $\bar{\alpha}$ către α arată că, pentru valorile mari ale lui n, $\bar{\alpha}$ diferă stochastic puțin de α și deci se justifică denumirea de funcție de estimație corectă dată lui $\bar{\alpha}$.

Există însă o infinitate de funcții $\bar{\alpha}$ care tind în probabilitate către α. Este deci natural să recurgem, pentru mai multă precizie, la reguli de convergență mai strînse. Convergența preferată în acest scop este cea în medie pătratică.

Dacă $\bar{\alpha}$ tinde în medie pătratică către α, cînd $n \to \infty$, avem evident o funcție de estimație corectă, deoarece convergența în medie pătratică implică convergența în probabilitate.

Putem avea în cazul convergenței în medie pătratică

$$M(\bar{\alpha}) = \alpha$$

sau

$$M(\bar{\alpha}) = \alpha + \theta(\alpha),$$

unde $\theta(\alpha)$ este o funcție de α.

In primul caz, funcția de estimație poartă numele de *absolut corectă.* In al doilea caz, $\theta(\alpha)$ reprezintă o eroare sistematică față de va-

oarea α, care tinde către zero cînd $n \to \infty$. Funcția de estimație este numai corectă, deoarece convergența în probabilitate a lui $\bar{\alpha}$ către α există.

Există o infinitate de funcții de estimație absolut corecte. De aceea, se pune din nou problema alegerii unei funcții de estimație absolut corecte mai precisă decît celelalte.

Să dăm un exemplu. Pentru aceasta, să plecăm de la un eveniment de probabilitate constantă p în fiecare probă. Supunem evenimentul la n probe independente și, cunoscînd rezultatele fiecărei probe, ne propunem pe baza lor să estimăm probabilitatea p, a cărei valoare numerică nu o cunoaștem.

Avem un singur parametru de estimat și anume p. Fie ν numărul de realizări ale evenimentului în cele n probe. Valoarea $\frac{\nu}{n}$ este o funcție de estimație absolut corectă a lui p, deoarece

$$M\left(\frac{\nu}{n}\right)=p, \quad D^2\left(\frac{\nu}{n}\right)=\frac{pq}{n}=O\left(\frac{1}{n}\right);$$

valoarea $\frac{\nu+1}{n+1}$ este o funcție de estimație corectă a lui α, deoarece

$$M\left(\frac{\nu+1}{n+1}\right)=p+\frac{1-p}{n+1}, \quad D^2\left(\frac{\nu+1}{n+1}\right)=\frac{npq}{(n+1)^2}=O\left(\frac{1}{n}\right).$$

Se vede că, deși pentru valorile mari ale lui n diferența dintre $\frac{\nu}{n}$ și $\frac{\nu+1}{n+1}$ este foarte mică, totuși este preferabil, din punct de vedere teoretic, să alegem pe $\frac{\nu}{n}$ ca funcție de estimație a lui p, în loc de $\frac{\nu+1}{n+1}$.

Să considerăm, referitor la același exemplu, o altă funcție de estimație absolut corectă.

Împărțim grupul de n experiențe efectuate în două: în primul grup punem primele n_1 experiențe efectuate, iar în al doilea cele n_2 experiențe care ne-au mai rămas ($n_1+n_2=n$).

Fie ν_1 numărul de bile albe în primele n_1 experiențe și ν_2 numărul de bile albe ieșite în cele n_2 experiențe ($\nu_1+\nu_2=\nu$).

Dacă k_1 și k_2 sînt două numere reale și pozitive date, alegem ca funcție de estimație a lui p expresia

$$\frac{k_1\nu_1+k_2\nu_2}{k_1n_1+k_2n_2}.$$

Efectuînd calculele, rezultă

$$M\left(\frac{k_1\nu_1+k_2\nu_2}{k_1n_1+k_2n_2}\right)=p,$$

$$D^2\left(\frac{k_1\nu_1+k_2\nu_2}{k_1n_1+k_2n_2}\right)=\frac{k_1n_1+k_2n_2}{(k_1n_1+k_2n_2)^2}=O\left(\frac{1}{n}\right).$$

Prin urmare,

$$\frac{k_1\nu_1+k_2\nu_2}{k_1n_1+k_2n_2}$$

reprezintă de asemenea o funcţie de estimaţie absolut corectă. Observăm însă că

$$D^2\left(\frac{k_1\nu_1+k_2\nu_2}{k_1n_1+k_2n_2}\right)=D^2\left(\frac{\nu}{n}\right)+\frac{(k_1-k_2)^2\,n_1n_2}{n\,(k_1n_1+k_2n_2)^2},$$

deci

$$D^2\left(\frac{k_1\nu_1+k_2\nu_2}{k_1n_1+k_2n_2}\right)\geq D^2\left(\frac{\nu}{n}\right),$$

egalitatea avînd loc numai cînd $k_1=k_2$ şi $\frac{k_1\nu_1+k_2\nu_2}{k_1n_1+k_2n_2}$ devine egal cu $\frac{\nu}{n}$.

Toate funcţiile de estimaţie de forma

$$\frac{k_1\nu_1+k_2\nu_2}{k_1n_1+k_2n_2}$$

distincte de $\frac{\nu}{n}$ au dispersia mai mare decît $\frac{\nu}{n}$. Deşi toate aceste funcţii de estimaţie sînt absolut corecte, totuşi cea preferabilă este aceea care dă naştere la cea mai mare concentrare a masei de probabilitate în jurul lui p, adică la cea mai mică dispersie. Prin urmare, vom alege pe $\frac{\nu}{n}$ ca funcţie de estimaţie a lui p, în locul expresiilor mai sus considerate.

Am văzut că dispersia lui $\frac{\nu}{n}$ este mai mică decît dispersia funcţiei de estimaţie absolut corecte

$$\frac{k_1\nu_1+k_2\nu_2}{k_1n_1+k_2n_2}.$$

Scanned with OKEN Scanner

Nu am demonstrat însă că dispersia lui $\frac{\nu}{n}$ este mai mică decît dispersia oricărei funcții de estimație absolut corecte.

Din cele de mai sus rezultă importanța practică a rezolvării acestei probleme, pe care o vom prezenta în paragraful următor.

Mai observăm că $\frac{\nu}{n}$ și $\frac{k_1\nu_1+k_2\nu_2}{k_1n_1+k_2n_2}$ nu numai că sînt convergente în medie pătratică, dar admit la limită distribuții normale. Astfel de funcții de estimație poartă numele de *funcții de estimație laplaciene*, după numele lui Laplace, care s-a ocupat cel dintîi cu legea normală de probabilitate.

Aceleași considerații se impun și în cazul în care ne propunem să estimăm o funcție $h(\alpha)$, în loc de α.

Vom spune că $\alpha(x)$ este o funcție de estimație corectă a lui $h(\alpha)$, dacă $\alpha(\vec{x})$ tinde în probabilitate către $h(\alpha)$, cînd $n\to\infty$.

In cazul convergenței în medie pătratică vom spune că $\alpha(\vec{x})$ este o funcție de estimație absolut corectă dacă avem în plus

$$M(\bar{\alpha})=h(\alpha).$$

Este evident că, dacă am găsit printr-o metodă oarecare o estimație pentru α, rezultă o estimație și pentru $h(\alpha)$. Dar nu întotdeauna estimația astfel obținută pentru $h(\alpha)$ este cea mai bună. De aceea este preferabil să se folosească, pentru estimarea lui (α), funcții de estimație speciale.

§ 4. Funcții de estimație eficiente

In exemplul de mai sus, printre funcțiile de estimație absolut corectă pe care le-am construit, există una avînd dispersia cea mai mică. Vom demonstra, referitor la aceasta, o teoremă generală privitoare la estimația parametrului dintr-o familie de distribuții statistice.

Fie $f(x,\alpha)$ *această familie și*

$$\bar{\alpha}(\vec{x})=\bar{\alpha}(x_1,x_2,\ldots,x_n)$$

o funcție dată de cele n observații. Vom arăta că *există întodeauna inegalitatea*

$$D^2[\bar{\alpha}(\vec{x})] \geq \frac{1}{n\int_{-\infty}^{+\infty}\left[\frac{\partial \ln f(x,\alpha)}{\partial\alpha}\right]^2 f(x,\alpha)\,dx}$$

și că egalitatea are loc numai în cazul cînd $f(x,\alpha)$ *are forma particulară*

$$f(x,\alpha)=-A'(\alpha)\,L(x)+A(\alpha)+N(x),$$

unde $L(x)$ *și* $N(x)$ *sînt funcții de* x, $A(\alpha)$ *o funcție de* α, *iar* $A'(\alpha)$ *derivata funcției* $A(\alpha)$.

Funcția $\bar{\alpha}(\vec{x})$ pentru care minimul este atins are expresia

$$\bar{\alpha}(\vec{x})=\frac{L(x_1)+L(x_2)+\ldots+L(x_n)}{n}.$$

Din relația

$$\int_{-\infty}^{+\infty} f(x,\alpha)\,dx=1$$

deducem, derivînd în raport cu α,

$$\int_{-\infty}^{+\infty}\frac{\partial f(x,\alpha)}{\partial\alpha}\,dx=0$$

sau

$$\int_{-\infty}^{+\infty}\frac{\partial \ln f(x,\alpha)}{\partial\alpha}f(x,\alpha)=0.$$

Prin urmare,

$$M\left(\frac{\partial \ln f(x,\alpha)}{\partial\alpha}\right)=0.$$

Prin ipoteză

$$M(\bar{\alpha})=\alpha.$$

Însă

$$M(\bar{\alpha})=\int\int\ldots\int_{V_n}\bar{\alpha}(x_1,x_2,\ldots,x_n)\prod_{i=1}^{n}f(x_i,\alpha)\,dx_1dx_2\ldots dx_n,$$

integrala fiind extinsă la întreg spațiul V_n în care se găsește punctul $P(x_1, x_2, \ldots, x_n)$.

Avem deci

$$\int\int\ldots\int_{V_n}\bar{\alpha}(x_1, x_2, \ldots, x_n)\prod_{i=1}^{n} f(x_i, \alpha)\,dx_1\,dx_2\ldots dx_n=\alpha.$$

Derivînd în raport cu α, obținem

$$\int\int\ldots\int_{V_n}\bar{\alpha}(x_1, x_2, \ldots, x_n)\sum_{i=1}^{n}\frac{\partial \ln f(x_i, \alpha)}{\partial\alpha}\prod_{i=1}^{n} f(x_i, \alpha)\,dx_1 dx_2 \ldots dx_n=1$$

sau, ținînd seama că valoarea medie a lui $\frac{\partial \ln f(x, \alpha)}{\partial\alpha}$ este nulă,

$$\int\int\ldots\int_{V_n}(\bar{\alpha}-\alpha)\sum_{i=1}^{n}\frac{\partial \ln f(x_i, \alpha)}{\partial\alpha}\prod_{i=1}^{n} f(x_i, \alpha)\,dx_1\,dx_2\ldots dx_n=1,$$

$$M\left[(\bar{\alpha}-\alpha)\sum_{i=1}^{n}\frac{\partial \ln f(x_i, \alpha)}{\partial\alpha}\right]=1.$$

Însă, potrivit inegalității lui Schwarz

$$\left\{M\left[(\bar{\alpha}-\alpha)\sum_{i=1}^{n}\frac{\partial \ln f(x_i, \alpha)}{\partial\alpha}\right]\right\}^2 \leqslant D(\bar{\alpha}-\alpha)\,D\left(\sum_{i=1}^{n}\frac{\partial \ln f(x_i, \alpha)}{\partial\alpha}\right);$$

$$1 \leqslant D(\bar{\alpha}-\alpha)\,D\left(\sum_{i=1}^{n}\frac{\partial \ln f(x_i, \alpha)}{\partial\alpha}\right).$$

Pe de altă parte, observațiile x_i fiind independente,

$$D^2\left(\sum_{i=1}^{n}\frac{\partial \ln f(x_i, \alpha)}{\partial\alpha}\right)=nD^2\left(\frac{\partial \ln f(x_i, \alpha)}{\partial\alpha}\right).$$

Rezultă

$$D^2(\bar{\alpha}-\alpha)\geqslant\frac{1}{nD^2\left(\frac{\partial \ln f(x_i, \alpha)}{\partial\alpha}\right)}$$

și deci prima parte a teoremei este demonstrată.

Pe de altă parte, ştim că inegalitatea lui Schwarz se transformă în egalitate numai dacă

$$K(\bar{\alpha}-\alpha)=\sum_{i=1}^{n}\frac{\partial \ln f(x_i,\ \alpha)}{\partial\alpha}, \qquad \text{(XVI.4,1)}$$

în cazul nostru K putînd fi o funcţie de parametrul α pentru orice valori x_i.

Egalitatea avînd loc pentru toate valorile lui $x_1, x_2, \ldots, x_n$, să facem $x_1=x$, $x_2=x_3=\ldots=x_n=0$.

Obţinem

$$\frac{\partial \ln f(x,\ \alpha)}{\partial\alpha}=F(\alpha)\,Q(x)+G(\alpha),$$

unde $F(\alpha)$, $G(\alpha)$ sînt funcţii de α, iar $Q(x)$ funcţie de x.

Rezultă atunci pentru α valoarea

$$\bar{\alpha}(\vec{x})=\frac{F(\alpha)}{K(\alpha)}\sum_{i=1}^{n}N(x_i)+\frac{nG(\alpha)}{K(\alpha)}+\alpha.$$

Deducem

$$F(\alpha)=h\,K(\alpha),$$
$$\frac{nG(\alpha)}{K(\alpha)}+\alpha=g,$$

h şi g fiind două constante.

Prin urmare,

$$\alpha(\vec{x})=h\sum_{i=1}^{n}Q(x_i)+g.$$

Notînd

$$L(x)=nh\,Q(x)+g,$$

obţinem

$$\bar{\alpha}(\vec{x})=\frac{\sum_{i=1}^{n}L(x_i)}{n}=\frac{\partial \ln f(x,\ \alpha)}{\partial\alpha}=\frac{K(\alpha)}{n}[L(x)-\alpha]$$

sau

$$\ln f(x,\ \alpha)=A'(\alpha)[L(x)-\alpha]+A(x)+N(x),$$

unde $L(x)$ şi $N(x)$ sînt funcţii numai de x, $A(\alpha)$ o funcţie de α, iar $A'(\alpha)$ derivata funcţiei $A(\alpha)$.

Să arătăm acum că, dacă $f(x,\ \alpha)$ are forma de mai sus, minimul indicat pentru dispersia lui $\bar{\alpha}$ este admis.

Scanned with OKEN Scanner

Deoarece $f(x, \alpha)$ este o densitate de probabilitate, trebuie să avem

$$\int_{-\infty}^{+\infty} e^{A'(\alpha)[L(x)-\alpha]+A(\alpha)+N(\alpha)}\,dx=1.$$

Derivînd în raport cu α, rezultă

$$A''(\alpha)\int_{-\infty}^{+\infty}[L(x)-\alpha]f(x,\alpha)\,dx=0 \qquad \text{(XVI.4,2)}$$

sau

$$M[L(x)]=\alpha$$

$$M[\overline{\alpha}(\vec{x})]=M\left(\frac{\sum_{i=1}^{n}L(x_i)}{n}\right)=\alpha.$$

Derivînd relația (XVI.4,1) în raport cu α, obținem

$$\int_{-\infty}^{+\infty}[L(x)-\alpha]^2 f(x,\alpha)\,dx=-\frac{1}{A''}$$

sau

$$D^2(L)=-\frac{1}{A''},$$

$$D^2(\overline{\alpha})=D^2\left(\frac{\sum_{i=1}^{n}L(x_i)}{n}\right)=-\frac{1}{nA''}.$$

Pe de altă parte

$$n\int_{-\infty}^{+\infty}\left(\frac{\partial \ln f(x,\alpha)}{\partial\alpha}\right)^2 f(x,\alpha)\,dx=nA''^2\int_{-\infty}^{+\infty}(L(x)-\alpha)^2 f(x,\alpha)=$$

$$=nA''^2D^2(L)=-A''n.$$

Prin urmare

$$D^2(\overline{\alpha})=\frac{1}{n\int_{-\infty}^{+\infty}\left[\frac{\partial \ln f(x,\alpha)}{\partial\alpha}\right]^2 f(x,\alpha)\,dx}. \qquad \text{(XVI.4,3)}$$

Teorema este demonstrată.

Fie $\alpha(\vec{x})$ o funcție de estimație absolut corectă pentru care minimul arătat este atins, adică avem îndeplinită egalitatea (XVI.4,2). In

acest caz, vom spune că $\overline{\alpha}(\vec{x})$ este o *funcţie de estimaţie eficientă.*

Fie acum $\overline{\alpha}(\vec{x})$ o funcţie de estimaţie absolut corectă oarecare. Raportul

$$e(\overline{\alpha}) = \frac{\dfrac{1}{n\int\limits_{-\infty}^{+\infty}\left[\dfrac{\partial \ln f(x,\ \alpha)}{\partial\alpha}\right]^2 f(x,\ \alpha)\,dx}}{D^2(\overline{\alpha})}$$

se numeşte *eficienţa* lui α. Este evident că $0 \leq e(\overline{\alpha}) \leq 1$. Pentru funcţiile de estimaţie eficiente, $e(\overline{\alpha}) = 1$.

Funcţii de estimaţie eficiente nu există decît în cazul particular cînd funcţia de probabilitate $f(x, \alpha)$ este de forma pe care am arătat-o. Pentru celelalte funcţii de estimaţie, cu cît $e(\overline{\alpha})$ este mai aproape de 1, cu atît precizia obţinută este mai mare.

Demonstraţia teoremei de mai sus a fost dată în cazul unei distribuţii continue $f(x, \alpha)$. Ea se păstrează şi pentru repartiţiile discontinue cu punctele de masă a_i $(i = 1, 2, \ldots)$ şi cu probabilitatea corespunzătoare $p_i(\alpha)$, unde α este parametrul de estimat, funcţiile $p_i(\alpha)$ îndeplinind condiţiile necesare pentru efectuarea calculelor intervenind în demonstraţie.

Să mai observăm că eficienţa $e(\overline{\alpha})$ este o funcţie de n. Dacă

$$\lim_{n\to\infty} e_n(\overline{\alpha}) = 1,$$

vom spune că funcţia de estimaţie este *eficient asimptotică.*

Teorema minimului dispersiei poate căpăta o demonstraţie directă folosind metodele clasice din calculul variaţiilor. Funcţia de estimaţie $\overline{\alpha}(x_1, x_2, \ldots, x_n)$ trebuie să îndeplinească relaţiile

$$\int\int\ldots\int_{V_n} \overline{\alpha}(x_1, x_2, \ldots, x_n)\, f(x_1, \alpha)\ldots f(x_n, \alpha)\, dx_1\, dx_2 \ldots dx_n = \alpha, \tag{XVI.4,4}$$

$$\int\int\ldots\int_{V_n} [\overline{\alpha}(x_1, x_2, \ldots, x_n) - \alpha]^2 f(x_1, \alpha)\ldots f(x_n, \alpha)\, dx_1\, dx_2 \ldots dx_n = \text{minim}, \tag{XVI.4,5}$$

integralele fiind extinse la întreg spaţiul V_n în care se găseşte punctul $P(x_1, x_2, \ldots, x_n)$.

Va trebui deci să determinăm funcția necunoscută $\bar{\alpha}(x_1, x_2, \ldots, x_n)$ prin condiția de minim (XVI.4,5), $\bar{\alpha}(x_1, \ldots, x_n)$ trebuind în același timp să verifice și relațiile (XVI.4,4).

Din relația (XVI.4,4), deducem, ca și în cazul precedent,

$$\int\int\ldots\int_{V_n} \bar{\alpha}(x_1, x_2, \ldots, x_n) \sum_{i=1}^{n} \frac{\partial \ln f(x_i, \alpha)}{\partial \alpha} \prod_{i=1}^{n} f(x_i, \alpha)\, dx_1 \ldots dx_n = 1,$$

admițînd că V_n este independent de α.

Din această relație, rezultă că minimul lui $\bar{\alpha}(\vec{x})$ este atins o dată cu minimul integralei

$$\int\int\ldots\int_{V_n} [\bar{\alpha}-\alpha]^2 \prod_{i=1}^{n} f(x_i, \alpha) - \frac{2}{k}\left[\bar{\alpha} \sum_{i=1}^{n} \frac{\partial \ln f(x_i, \alpha)}{\partial \alpha} \prod_{i=1}^{n} f(x_i, \alpha)\right] \times$$

$$\times\, dx_1 \ldots dx_n,$$

unde k este o constantă oarecare față de variabilele $x_1, x_2, \ldots, x_n$, dar poate să depindă de α.

Sîntem în prezența unei probleme Euler din calculul variațiilor.

O valoare extremală rezultă din ecuația

$$\frac{\partial}{\partial \alpha}\left\{(\bar{\alpha}-\alpha)^2 \prod_{i=1}^{n} f(x_i, \alpha) - \frac{2}{k}\left[\bar{\alpha} \sum_{i=1}^{n} \frac{\partial \ln f(x_i, \alpha)}{\partial x_i} \prod_{i=1}^{n} f(x_i, \alpha)\right]\right\} = 0$$

sau

$$(\bar{\alpha}-\alpha) - \frac{1}{k} \sum_{i=1}^{n} \frac{\partial \ln f(x_i, \alpha)}{\partial x_i} = 0,$$

adică tocmai relația (XVI.4,1). Restul demonstrației urmează ca mai sus.

Am dat această demonstrație numai ca o indicație, deoarece nu este suficient de riguroasă, atît în ce privește stabilirea condițiilor în care ecuația lui Euler este aplicabilă, cît și în privința faptului că valoarea extremă găsită reprezintă un minim. Totuși, metoda trebuie să fie menționată pentru că în numeroase împrejurări concrete ea conduce la acest minim și atunci reprezintă calea cea mai naturală pentru rezolvarea problemei.

Scanned with OKEN Scanner

Să considerăm, înainte de a încheia acest paragraf, două funcţii de extimaţie eficiente $\overline{\alpha}_1(x_1, \ldots, x_n)$ şi $\overline{\alpha}_2(x_1, \ldots, x_n)$.

Se vede uşor că şi *funcţia* $\alpha(x_1, x_2, \ldots, x_n)$ *definită prin relaţia*

$$\overline{\alpha}(x_1, x_2, \ldots, x_n) = \lambda\alpha_1(x_1, x_2, \ldots, x_n) + (1-\lambda)\,\overline{\alpha}_2(x_1, \ldots, x_n),$$

unde λ *este o constantă oarecare, este o funcţie de estimaţie.*

Deoarece $\overline{\alpha}_1(x)$ şi $\overline{\alpha}_2(x)$ sînt amîndouă funcţii de estimaţie eficiente, avem

$$D^2(\overline{\alpha}_1) = D^2(\overline{\alpha}_2). \qquad \text{(XVI.4,6)}$$

Pe de altă parte,

$$D^2(\overline{\alpha}) = M[\lambda(\overline{\alpha}_1 - \alpha) + (1-\lambda)(\overline{\alpha}_2 - \alpha)]^2 = \lambda^2 D^2(\overline{\alpha}_1) +$$

$$+ 2\lambda(1-\lambda)\,M(\overline{\alpha}_1 - \alpha)(\overline{\alpha}_2 - \alpha) + (1-\lambda)^2 D^2(\alpha_2).$$

Insemnînd prin ρ coeficientul de corelaţie al funcţiilor $\overline{\alpha}_1$ şi $\overline{\alpha}_2$, avem, prin definiţie,

$$\rho = \frac{M(\overline{\alpha}_1 - \alpha)(\overline{\alpha}_2 - \alpha)}{D(\overline{\alpha}_1)\,D(\overline{\alpha}_2)}.$$

De aici rezultă că putem scrie, ţinînd seamă de (XVI.4,6),

$$D^2(\overline{\alpha}) = D^2(\overline{\alpha}_1)[\lambda^2 + 2\lambda(1-\lambda)\rho + (1-\lambda)^2].$$

Sau

$$2\lambda(1-\lambda)(\rho-1) = \frac{D^2(\overline{\alpha})}{D^2(\overline{\alpha}_1)}.$$

Avem însă

$$\frac{D^2(\overline{\alpha})}{D^2(\overline{\alpha}_1)} = 1,$$

deoarece $\overline{\alpha}_1$ este o funcţie de estimaţie eficientă.

Prin urmare, oricare ar fi λ, trebuie să avem

$$2\lambda(1-\lambda)(\rho-1) \leqslant 1,$$

ceea ce nu este posibil decît dacă $\rho = 1$.

Prin urmare, două funcţii de estimaţie eficiente au nu numai dispersiile egale, dar şi coeficientul de corelaţie egal cu unu, deci ele sînt egale, exceptînd punctele de probabilitate nulă.

Scanned with OKEN Scanner

§ 5. Funcții de estimație suficiente

Dăm acum o nouă demonstrație teoremei fundamentale din teoria estimației cu privire la minimul variației unei estimații absolut corecte. Demonstrația pe care am dat-o are avantajul simplicității. Demonstrația care urmează, datorită lui D. Dugué, pune în evidență un fapt nou și anume legătura dintre funcțiile de estimație eficiente și existența unor statistici exhaustive. Cu această ocazie, vom arăta că teorema minimului dispersiei se poate extinde la funcțiile de estimație corecte.

Vom pleca tot de la o distribuție continuă cu funcția de frecvență $f(x, \alpha)$.

Fie $\bar{\alpha}(x_1, x_2, \ldots, x_n) = \alpha(\vec{x})$ o funcție de estimație corectă. Atunci

$$M(\alpha(\vec{\alpha})) = \alpha + \theta(\alpha),$$

$\theta(\alpha)$ fiind o funcție de α și de n, tinzînd către zero cînd $n \to \infty$.

Ecuația $\bar{\alpha}(\vec{x}) = 0$ reprezintă în R_n o hipersuprafață S. Punctul $P(x_1, x_2, \ldots, x_n)$ se găsește pe această hipersuprafață. Vom alege pe S coordonatele $\xi_1, \xi_2, \ldots, \xi_{n-1}$ în așa fel încît să avem o corespondență biunivocă între sistemele de coordonate $(x_1, x_2, \ldots, x_n)$ și $(\bar{\alpha}, \xi_1, \xi_2, \ldots, \xi_n)$ referitoare la spațiul R_n.

Masa de probabilitate concentrată în punctul P, raportată la sistemul $(x_1, x_2, \ldots, x_n)$, este

$$f(x_1, \alpha) f(x_2, \alpha) \ldots f(x_n, \alpha),$$

iar la sistemul

$$(\alpha', \xi_1, \xi_2, \ldots, \xi_{n-1}),$$

este

$$g(\bar{\alpha}, \alpha)\, h(\xi_1, \xi_2, \ldots, \xi_{n-1} \mid \bar{\alpha}, \alpha)\, d\bar{\alpha}\, d\xi_1\, d\xi_2 \ldots d\xi_{n-1},$$

unde $g(\bar{\alpha}, \alpha)$ reprezintă legea de probabilitate a lui $\bar{\alpha}$, iar $h(\xi_1, \xi_2, \ldots, \xi_{n-1} \mid \bar{\alpha}, \alpha)$ legea de probabilitate a punctului P pe hipersuprafața S. Se știe că

$$f(x_1, \alpha) f(x_2, \alpha) \ldots f(x_n, \alpha) \left| \frac{D(x_1, x_2, \ldots, x_n)}{D(\bar{\alpha}, \xi_1, \ldots, \xi_{n-1})} \right| =$$

$$= g(\bar{\alpha}, \alpha)\, h(\xi_1, \xi_2, \ldots, \xi_{n-1} \mid \bar{\alpha}, \alpha). \qquad \text{(XVI.5,1)}$$

Admitem că există derivatele $\frac{\partial f}{\partial \alpha}$, $\frac{\partial g}{\partial \alpha}$, $\frac{\partial h}{\partial \alpha}$ şi că

$$\frac{\partial f}{\partial \alpha} < F_0(x), \quad \left|\frac{\partial g}{\partial \alpha}\right| < G_0(\bar{\alpha}), \quad \frac{\partial h}{\partial \alpha} < H_0(\xi_1, \xi_2, \ldots, \xi_{n-1}; \bar{\alpha})$$

pentru orice α, funcţiile $F_0(x)$, $G_0(\bar{\alpha})$ şi $H_0(\xi_1, \xi_2, \ldots, \xi_{n-1}; \bar{\alpha})$ fiind integrabile în R_n.
Din relaţia

$$\int_{-\infty}^{+\infty} f(x, \alpha)\, dx = 1$$

deducem, derivînd în raport cu α,

$$\int_{-\infty}^{+\infty} \frac{\partial \ln f(x, \alpha)}{\partial \alpha} f(x, \alpha)\, dx = 0$$

şi deci pe hipersuprafaţa $\bar{\alpha}$

$$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \ldots \int_{-\infty}^{+\infty} h(\xi_1, \xi_2, \ldots, \xi_{n-1} \mid \bar{\alpha}, \alpha)\, d\xi_1 d\xi_2 \ldots d\xi_{n-1} = 1.$$

Derivînd în raport cu α, deducem în acelaşi mod

$$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \ldots \int_{-\infty}^{+\infty} \frac{\partial \ln h}{\partial \alpha} h(\xi_1, \xi_2, \ldots, \xi_{n-1} \mid \bar{\alpha}, \alpha)\, d\xi_1 d\xi_2 \ldots d\xi_{n-1} = 0.$$

Derivînd relaţia (XVI.5,1) în raport cu α şi ţinînd seama că jacobianul

$$\frac{D(x_1, x_2, \ldots, x_n)}{D(\bar{\alpha}, \xi_1, \xi_2, \ldots, \xi_{n-1})}$$

nu depinde de α, rezultă că

$$\sum_{i=1}^{n} \frac{\partial \ln f(x_i, \alpha)}{\partial \alpha} = \frac{\partial \ln g}{\partial \alpha} + \frac{\partial \ln h}{\partial \alpha}.$$

Ridicînd la pătrat în ambii membri ai acestei relații, integrînd în întreg spațiul și ținînd seama de egalitățile pe care le-am stabilit, rezultă,

$$n\int_{-\infty}^{+\infty}\left[\frac{\partial \ln f(x,\alpha)}{\partial\alpha}\right]^2 f(x,\alpha)\,dx=\int_{-\infty}^{+\infty}\left(\frac{\partial \ln g}{\partial\alpha}\right)^2 g(\bar{\alpha},\alpha)\,d\bar{\alpha}+$$

$$+\int_{-\infty}^{+\infty} g(\bar{\alpha},\alpha)\,d\alpha\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\dots\int_{-\infty}^{+\infty}\left(\frac{\partial \ln h}{\partial\alpha}\right)h\,d\xi_1 d\xi_2\dots d\xi_{n-1} \qquad \text{(XVI.5,2)}$$

sau

$$n\int_{-\infty}^{+\infty}\left[\frac{\partial \ln f(x,\alpha)}{\partial\alpha}\right]^2 f(x,\alpha)\,dx\geqq\int_{-\infty}^{+\infty}\left(\frac{\partial \ln g(\bar{\alpha},\alpha)}{\partial\alpha}\right)^2 g(\bar{\alpha},\alpha)\,d\alpha. \qquad \text{(XVI.5,3)}$$

Pe de altă parte,

$$\int_{-\infty}^{+\infty} g(\bar{\alpha},\alpha)\,d\bar{\alpha}=1,\quad \int_{-\infty}^{+\infty}\bar{\alpha}\,g(\bar{\alpha},\alpha)\,d\alpha=\alpha+\theta(\alpha).$$

Derivînd în raport cu α, obținem

$$\int_{-\infty}^{+\infty}\frac{\partial \ln g(\bar{\alpha},\alpha)}{\partial\alpha}\,g(\bar{\alpha},\alpha)\,d\alpha=0$$

$$\int_{-\infty}^{+\infty}\bar{\alpha}\,\frac{\partial \ln g(\bar{\alpha},\alpha)}{\partial\alpha}\,g(\bar{\alpha},\alpha)\,d\alpha=1+\theta(\alpha).$$

Prin urmare,

$$\int_{-\infty}^{+\infty}(\bar{\alpha}-\alpha)\,\frac{\partial \ln g(\bar{\alpha},\alpha)}{\partial\alpha}\,g(\bar{\alpha},\alpha)\,d\alpha=1+\theta'(\alpha).$$

Însă, potrivit inegalității lui Schwarz,

$$\left[\int_{-\infty}^{+\infty}(\bar{\alpha}-\alpha)\frac{\partial \ln g}{\partial\alpha}\,g\,d\alpha\right]^2=\left[\int_{-\infty}^{+\infty}(\bar{\alpha}-\alpha)\sqrt{g}\,\frac{\partial \ln g}{\partial\alpha}\sqrt{g}\,d\bar{\alpha}\right]^2\leqq$$

$$\leqq\int_{-\infty}^{+\infty}(\bar{\alpha}-\alpha)^2 g\,d\bar{\alpha}\int_{-\infty}^{+\infty}\left(\frac{\partial \ln g}{\partial\alpha}\right)^2 g\,d\bar{\alpha},$$

sau

$$(1+\theta')^2 \leqslant M(\bar{\alpha}-\alpha)^2 \int_{-\infty}^{+\infty} \left(\frac{\partial \ln g}{\partial \alpha}\right)^2 g\, d\bar{\alpha}.$$

Comparînd această inegalitate cu (XVI.5,1), rezultă că

$$M(\bar{\alpha}-\alpha)^2 \geqslant \frac{(1+\theta')^2}{n \int_{-\infty}^{+\infty} \left(\frac{\partial \ln f(x,\alpha)}{\partial \alpha}\right)^2 f(x,\alpha)\, dx}.$$

Se obţine deci un minim şi pentru cazul funcţiilor de estimaţie corecte. Trebuie observat că de data aceasta minimul se referă nu la variaţia lui $\bar{\alpha}$, ci la valoarea medie a pătratului abaterii lui α faţă de $\bar{\alpha}$. De altfel, pentru estimaţiile curente ne interesează această medie, nu dispersia, deoarece problema noastră constă în estimaţia parametrului α.

In cazul funcţiilor de estimaţie absolut corecte, $\theta'=0$ şi teorema se reduce la precedenta.

Din inegalitatea (XVI.5,2) rezultă că minimul poate fi atins numai dacă

$$\frac{\partial \ln h(\xi_1, \xi_2, \ldots, \xi_{n-1} | \bar{\alpha}, \alpha)}{\partial \alpha} = 0,$$

ceea ce înseamnă că $h(\xi, \bar{\alpha})$ nu conţine pe α.

Din relaţia (XVI.5,1) deducem că, pentru estimaţia lui α, numai $g(\bar{\alpha}, \alpha)$ poate să ne dea o informaţie, deoarece $h(\xi_1, \xi_2, \ldots, \xi_{n-1} | \bar{\alpha})$ nu conţine pe α. Prin urmare, $\bar{\alpha}(x_1, x_2, \ldots, x_n)$ este o statistică exhaustivă pentru α, referitoare la grupul de informaţie $(x_1, x_2, \ldots, x_n)$. Am demonstrat astfel proprietatea importantă pentru aplicaţii că orice funcţie de estimaţie eficientă $\bar{\alpha}(x_1, x_2, \ldots, x_n)$ reprezintă o statistică exhaustivă.

Nu însă orice statistică exhaustivă dă naştere la o funcţie de estimaţie eficientă, după cum vom arăta cu exemple.

Funcţia $\bar{\alpha}(x_1, x_2, \ldots, x_n)$, care în acelaşi timp este o statistică exhaustivă pentru estimaţia parmetrului α în raport cu grupul de informaţie $(x_1, x_2, \ldots, x_n)$, poartă numele de *funcţie de estimaţie suficientă*. Din definiţie, rezultă că orice funcţie de estimaţie eficientă este şi suficientă.

Să examinăm legătura dintre funcția de estimație eficientă și funcția de estimație suficientă. Am văzut că, pentru ca să existe o funcție de estimație eficientă, trebuie ca

$$\ln f(x, \alpha) = A'(\alpha)[L(x) - \alpha] + A(\alpha) + N(x).$$

Să considerăm o familie de distribuții $f(x, \alpha)$, care nu este de forma de mai sus, adică pentru care nu există o funcție de estimație eficientă. Ne putem întreba dacă efectuînd o schimbare $\alpha = h(\beta)$ asupra parametrului α, familia de densități $f[x, h(\beta)]$, care depinde acum de parmetrul β, poate să admită o funcție de estimație eficientă. Întrebarea este firească, deoarece estimația lui β atrage după sine și estimația lui α.

Cu notațiile întrebuințate, va trebui să avem

$$\ln f[x, h(\beta)] = A'(\beta)[L(x) - \beta] + A(\beta) + N(x).$$

Din $\alpha = h(\beta)$ rezultă că $\beta = h^{-1}(\alpha)$; făcînd această înlocuire în relația de mai sus, obținem

$$\ln f(x, \alpha) = A_1(\alpha)\psi_1(x) + A_0(\alpha) + \psi_0(x),$$

unde

$$A_1(\alpha) = A(\beta), \quad \psi_1(x) = L(x),$$
$$A_0(\alpha) = A(\beta) - \beta A'(\beta), \quad \psi_0(x) = N(x).$$

După cum vom vedea la punctul următor, forma găsită pentru $f(x, \alpha)$ arată existența unei statistici exhaustive.

Din relațiile precedente rezultă imediat

$$\beta = -\frac{A_0'(\alpha)}{A_1'(\alpha)}.$$

Am demonstrat astfel că singurele familii de densități depinzînd de parametrul α, care prin transformarea $\beta = h(\alpha)$ dau naștere la o familie depinzînd de β, care admite o funcție de estimație eficientă, sînt familiile de densități care admit o statistică exhaustivă.

Ajungem astfel la o funcție de estimație eficientă a lui β, dată de relația

$$\bar{\alpha}(\vec{x}) = \frac{\sum_{i=1}^{n} L(x_i)}{n},$$

care constituie statistica exhaustivă a familiei.

Din relaţia $\beta = h(\alpha)$ rezultă o funcţie de estimaţie a lui α, care însă nu mai este eficientă şi pe care o numim *suficientă*. Valoarea $\bar{\alpha}(\vec{x})$ furnizează deci o funcţie de estimaţie eficientă pentru β şi o funcţie de estimaţie suficientă pentru α.

§ 6. Statistici exhaustive

Ne ocupăm aici îndeaproape de noţiunea de statistică exhaustivă, deoarece rezultatele obţinute pînă acum nu pun în adevărata lumină acest instrument de calcul important din teoria estimaţiei.
Fie

$$\bar{\alpha}(x_1, x_2, \ldots, x_n) = c_i \qquad (i = 1, 2, \ldots, m;\ m < n)$$

m hipersuprafeţe din spaţiul R_n.

Densitatea de repartiţie a punctului $P(x_1, x_2, \ldots, x_n)$ este $\prod_{i=1}^{n} f(x_i, \alpha)$.
Dacă probabilitatea condiţionată a punctului P, după ce ştim că avem

$$\bar{\alpha}(\vec{x}) = c_i \qquad (i = 1, 2, \ldots, m),$$

nu depinde de α şi $m < n$ este cel mai mic număr pentru care această probabilitate este adevărată, vom spune că funcţiile $\bar{\alpha}_1(\vec{x})$ constituie o *statistică exhaustivă de ordinul m.*
Hipersuprafeţele $\bar{\alpha}_i = c_i$ $(i = 1, 2, \ldots, m)$ formează în R_n o familie Σ de varietăţi $(n-m)$-dimensionale, depinzînd de parametrii c_i. Prin fiecare punct $P_0(x_1^0, x_2^0, \ldots, x_n^0)$ trece o astfel de varietate Σ_0 corespunzătoare valorilor parametrilor $(c_1^0, c_2^0, \ldots, c_n^0)$ deduşi din relaţiile

$$\bar{\alpha}_i(x_1^0, x_2^0, \ldots, x_n^0) = c_i^0 \qquad (i = 1, 2, \ldots, m).$$

Probabilitatea poziţiei pe care punctul P_0 o ocupă pe Σ_0 este independentă de α. Prin urmare, poziţia lui P_0 pe Σ_0 nu ne poate procura nici o informaţie pentru determinarea lui α. Singura informaţie posibilă pentru estimaţia lui α în raport cu observaţiile $(x_1, x_2, \ldots, x_n)$ este dată de funcţiile $\alpha_i(\vec{x})$ $(i = 1, 2, \ldots, m)$.
In loc să plecăm de la datele iniţiale $(x_1, x_2, \ldots, x_n)$ rezultate din observaţie, putem folosi pentru scopul nostru expresiile $\alpha_1(\vec{x}), \alpha_2(\vec{x}), \ldots$ $\ldots, \alpha_n(\vec{x})$ $(m < n)$ care conţin toate elementele care ne pot furniza

informațiile asupra lui α. Deoarece $m < n$, obținem o reducere a numărului de elemente cu care lucrăm. În baza acestei proprietăți, mulțimea funcțiilor $\alpha_i(\vec{x})$ poartă numele de *rezumat exhaustiv* pentru estimarea lui α.

Preferăm totuși denumirea de statistică exhaustivă, care corespunde operațiilor pe care le efectuăm. După ce am făcut un număr n de observații x_i, prelucrarea statistică pe care trebuie s-o facem în vederea estimării lui α constă în calcularea funcțiilor $\alpha_i(\vec{x})$ care reprezintă singurul material statistic în virtutea căruia putem trage concluzii asupra valorii lui α.

Înainte de a trece la determinarea funcțiilor de probabilitate $f(x, \alpha)$ pentru care este posibilă existența unei statistici exhaustive, vom face unele observații necesare pentru precizarea acestei noțiuni.

Astfel, dacă funcțiile $\alpha_i(\vec{x})$ constituie o statistică exhaustivă de ordinul m, orice sistem $\beta_i(\alpha_1, \alpha_2, \ldots, \alpha_m)$ de m funcții independente de α_i reprezintă, de asemenea, o statistică exhaustivă de ordinul m. Această statistică exhaustivă nu este însă distinctă de precedenta, pentru că dă naștere la aceeași familie de varietăți Σ.

De asemenea, existența unei statistici exhaustive implică legi particulare de probabilitate pentru distribuțiile pe care le considerăm. Aceste legi particulare de probabilitate se întîlnesc însă la majoritatea distribuțiilor clasice din statistica matematică, justificînd interesul pe care-l avem pentru a studia această metodă particulară de cercetare.

Să vedem în ce condiții există o statistică exhaustivă de ordinul minim m. Pentru aceasta, conform definiției, este necesar ca relația (XVI.5,1) din paragraful precedent să devină

$$f(x_1, \alpha) f(x_2, \alpha) \ldots f(x_n, \alpha) \left| \frac{D(x_1, x_2, \ldots, x_n)}{D(\alpha_1, \alpha_2, \ldots, \alpha_m, \xi_1, \ldots, \xi_{n-m})} \right| =$$

$$= g(\alpha_1, \alpha_2, \ldots, \alpha_m \mid \alpha)\, h(\xi_1, \xi_2, \ldots, \xi_{n-m} \mid \alpha_1, \alpha_2, \ldots, \alpha_m), \quad \text{(XVI.6,1)}$$

funcția $h(\xi_1, \xi_2, \ldots, \xi_{n-m} \mid \alpha_1, \alpha_2, \ldots, \alpha_m)$ fiind independentă de α.

De asemenea, determinantul funcțional $\dfrac{D(x_1, x_2, \ldots, x_n)}{D(\alpha_1, \alpha_2, \ldots, \alpha_m, \xi_1, \ldots, \xi_{n-m})}$ este independent de α.

Scanned with OKEN Scanner

Derivînd relaţia (XVI.6,1) în raport cu α şi notînd pentru simplificare

$$\frac{\partial \ln f(x, \alpha)}{\partial \alpha} = \varphi(x, \alpha),$$

$$\frac{\partial \ln g(\alpha_1, \alpha_2, \ldots, \alpha_m \mid \alpha)}{\partial \alpha} = G(\alpha_1, \alpha_2, \ldots, \alpha_m, \alpha),$$

$$\alpha_k(x_1, x_2, \ldots, x_n) = y_k(x_1, x_2, \ldots, x_n) \quad (k=1, 2, \ldots, n),$$

obţinem relaţia

$$\sum_{i=1}^{n} \varphi(x_i, \alpha) = G(y_1, y_2, \ldots, y_m, \alpha). \qquad \text{(XVI.6,2)}$$

Derivînd în raport cu x_i, obţinem

$$\frac{\partial \varphi(x_i, \alpha)}{\partial x_i} = \sum_{k=1}^{n} \frac{\partial y_k}{\partial x_i} \cdot \frac{\partial G(y, \alpha)}{\partial y_k}. \qquad \text{(XVI.6,3)}$$

Să dăm lui α succesiv valorile particulare $\beta_1, \beta_2, \ldots, \beta_m$. Egalitatea de mai sus se scrie

$$\frac{\partial \varphi(x_i, \beta_s)}{\partial x_i} = \sum_{k=1}^{n} \frac{\partial y_k}{\partial x_i} \cdot \frac{\partial G(y, \beta)}{\partial y_k} \qquad (s=1, 2, \ldots, m). \qquad \text{(XVI.6,4)}$$

Eliminînd expresiile $\frac{\partial y_k}{\partial x_i}$ între relaţiile (XVI.6,3) şi (XVI.6,4), obţinem

$$\begin{vmatrix} \frac{\partial \varphi(x_i, \alpha)}{\partial x_i} & \frac{\partial \varphi(x_i, \beta_1)}{\partial x_i} & \ldots & \frac{\partial \varphi(x_i, \beta_m)}{\partial x_i} \\ \frac{\partial G(y, \alpha)}{\partial y_1} & \frac{\partial G(y, \beta_1)}{\partial y_1} & \ldots & \frac{\partial G(y, \beta_m)}{\partial y_1} \\ \ldots & \ldots & \ldots & \ldots \\ \frac{\partial G(y, \alpha)}{\partial y_m} & \frac{\partial G(y, \beta_1)}{\partial y_m} & \ldots & \frac{\partial G(y, \beta_m)}{\partial y_m} \end{vmatrix}.$$

Minorii elementelor din prima linie nu pot fi identic nuli, deoarece în acest caz statistica exhaustivă ar înceta să fie de ordinul minim m.

Folosind această proprietate și dînd lui x_i valorile $x_1, x_2, \ldots, x_{m+1}$ rezultă că trebuie să avem

$$\begin{vmatrix} \frac{\partial\varphi(x_1,\alpha)}{\partial x_1} & \frac{\partial\varphi(x_1,\beta_1)}{\partial x_1} & \ldots & \frac{\partial\varphi(x_1,\beta_m)}{\partial x_1} \\ \frac{\partial\varphi(x_2,\alpha)}{\partial x_2} & \frac{\partial\varphi(x_2,\beta_1)}{\partial x_2} & \ldots & \frac{\partial\varphi(x_2,\beta_m)}{\partial x_2} \\ \ldots\ldots & \ldots\ldots & \ldots & \ldots\ldots \\ \frac{\partial\varphi(x_{m+1},\alpha)}{\partial x_{m+1}} & \frac{\partial\varphi(x_{m+1},\beta_1)}{\partial x_{m+1}} & \ldots & \frac{\partial\varphi(x_{m+1},\beta_m)}{\partial x_{m+1}} \end{vmatrix} = 0.$$

Dezvoltînd acest determinant, obținem

$$\frac{\partial\varphi(x_1,\alpha)}{\partial x_1}\Phi(x_2, x_3, \ldots, x_{m+1};\ \beta_1, \beta_2, \ldots, \beta_m) +$$
$$+\frac{\partial\varphi(x_2,\alpha)}{\partial x_2}\Phi(x_3, \ldots, x_{m+1};\ \beta_1, \ldots, \beta_m) + \ldots$$
$$+\frac{\partial\varphi(x_{m+1},\alpha)}{\partial x_{m+1}}\Phi(x_1, x_2, \ldots, x_m;\ \beta_1, \beta_2, \ldots, \beta_m) = 0.$$

Să notăm $x_1 = x$ și să dăm variabilelor $x_2, x_3, \ldots, x_{m+1}$ valorile constante $a_1, a_2, \ldots, a_m$ respectiv. Din relația precedentă se vede că $\frac{\partial\varphi(x_1,\alpha)}{\partial x}$ este de forma

$$\frac{\partial\varphi(x,\alpha)}{\partial\alpha} = \sum_{k=1}^{m} H_k(\alpha)\, Q_k(x),$$

unde $H_k(\alpha)$ și $Q_k(x)$ sînt respectiv funcții de α și de x.

Ținînd seama de faptul că

$$\varphi(x,\alpha) = \frac{\partial \ln f(x,\alpha)}{\partial\alpha},$$

prin integrări succesive rezultă

$$\ln f(x,\alpha) = \sum_{k=1}^{m} A_k(\alpha)\,\Psi_k(x) + A_0(\alpha) + \Psi_0(x). \qquad \text{(XVI.6,5)}$$

Funcțiile $A_k(\alpha)$ și $\Psi_k(x)$ $(k = 0, 1, 2, \ldots, m)$ sînt funcții arbitrare de α și de x respectiv, satisfăcînd relația

$$\int_{-\infty}^{+\infty} f(x,\alpha)\, d\alpha = 1.$$

In plus, funcţiile $A_k(\alpha)$ trebuie să fie liniar independente. Aceeaşi condiţie trebuie s-o îndeplinească şi funcţiile $\Psi_k(x)$ pentru că altfel statistica exhaustivă ar fi de un ordin minim mai mic decît m.

Se vede că forma dată de relaţia (XVI.6,5) pentru $f(x,\alpha)$ este necesară cu condiţia ca derivatele de primele două ordine în raport cu α şi cu x, care intră în aceste calcule, să existe.

Vom arăta acum că relaţia (XVI.6,5) este suficientă pentru existenţa unei statistici exhaustive de ordinul minim m. Intr-adevăr, din egalitatea

$$f(x,\alpha)=e^{\sum_{k=1}^{n} A_k(\alpha)\Psi_k(x)+A_0(\alpha)+\Psi_0(x)},$$

deducem

$$\prod_{i=1}^{n} f(x_i,\alpha)=e^{n\sum_{i=1}^{n}\Psi_0(x_i)}\,e^{\sum_{i=1}^{n}A_h(\alpha)y_k(x)+nA_0(x)},$$

unde

$$y_k(x)=\sum_{i=1}^{n}\Psi_k(x_i).$$

Relaţia precedentă ne arată că, într-adevăr, se obţine o statistică exhaustivă de ordin minim m şi că funcţiile $\alpha_i(x_1,x_2,\ldots,x_n)$ care ne conduc la această statistică sînt egale, după cum rezultă din relaţia de mai sus, cu

$$\alpha_k(\vec{x})=\sum_{i=1}^{n}\Psi_k(x_i). \qquad \text{(XVI.6,6)}$$

In cazul particular, în care statistica exhaustivă este de ordinul întîi, avem

$$\ln f(x,\alpha)=A_i(\alpha)\Psi_1(x)+A_0(\alpha)+\Psi_0(x). \qquad \text{(XVI.6,7)}$$

Statistica exhaustivă este dată de funcţia

$$\overline{\alpha}_1(\vec{x})=\sum_{i=1}^{n}\Psi(x_i).$$

Pe de altă parte, am arătat că pentru ca să existe o estimație eficientă, este necesar și suficient ca

$$\ln f(x,\alpha)=-A'(\alpha)\,[L(x)-\alpha]+A(\alpha)+N(x), \qquad \text{(XVI.6,8)}$$

estimația eficientă fiind dată de funcția

$$\overline{\alpha}_1(\vec{x})=\frac{\sum_{i=1}^{n}L(x_i)}{n}.$$

Comparînd relațiile (XVI.6,7) și (XVI.6,8) se vede imediat că, pentru a avea o funcție de estimație eficientă, este necesar să existe o statistică exhaustivă de ordin minim unu. Reciproca nu este însă adevărată.

In acest caz, funcția care dă o statistică exhaustivă reprezintă și o funcție de estimație a parametrului α, ceea ce nu se întîmplă cînd avem o statistică exhaustivă de ordin mai mare decît unu.

§ 7. Metoda verosimilității maxime

Să considerăm densitatea de repartiție

$$P(x,\alpha)=\prod_{i=1}^{n}f(x_i,\alpha),$$

corespunzătoare punctului $(x_1,x_2,\ldots,x_n)$ din R_n.

Valorile $(x_1,x_2,\ldots,x_n)$ fiind considerate ca date, ele fiind rezultatul experienței, vom considera ca valoare cea mai verosimilă pentru α valoarea pentru care probabilitatea $P(x,\alpha)$ devine maximă.

Fie α^* o valoare a lui α pentru care probabilitatea $P(x,\alpha)$ este maximă. In general α^* depinde de $x_1,x_2,\ldots,x_n$, adică

$$\alpha^*=\alpha^*(x_1,x_2,\ldots,x_n)=\alpha^*(\vec{x}).$$

Vom elimina valorile lui α^* care nu depind de $x_1,x_2,\ldots,x_n$ și vom considera ca estimații ale lui α numai valorile lui α^* care sînt funcții de $x_1,x_2,\ldots,x_n$. Maximul lui $P(x,\alpha)$ are loc în același timp cu maximul lui $\ln P(x,\alpha)$. Funcția de estimație de verosimilitate maximă $\alpha^*(\vec{x})$ trebuie deci să verifice ecuația

$$\frac{\partial \ln P(x,\alpha)}{\partial\alpha}=\sum_{i=1}^{n}\frac{\partial \ln f(x_i,\alpha)}{\partial\alpha}=0 \quad (\alpha=\alpha^*). \qquad \text{(XVI.7,1)}$$

Să arătăm că, dacă există o funcție de estimație eficientă $\bar{\alpha}(\vec{x})$ ea este dată de funcția de estimație de verosimilitate maximă $\alpha^*(\vec{x})$. Intr-adevăr, dacă există o funcție de estimație eficientă $\ln f(x, \alpha)$, ea trebuie să fie de forma

$$\ln f(x, \alpha) = A'(\alpha)\,[L(x) - \alpha] + A(\alpha) + N(x),$$

$$\bar{\alpha}(\vec{x}) = \frac{\sum_{i=1}^{n} L(x_i)}{n}.$$

Din ecuația verosimilității maxime (XVI.7,1) rezultă

$$A''(\alpha^*) \sum_{i=1}^{n} [L(x_i) - \alpha^*] = 0.$$

Singura soluție, funcție de $x_1, x_2, \ldots, x_n$, este

$$\alpha^*(\vec{x}) = \frac{\sum_{i=1}^{n} L(x_i)}{n},$$

care coincide cu funcția de estimație eficientă $\bar{\alpha}(x)$.
Să presupunem acum că există o funcție de estimație suficientă. In acest caz am arătat că

$$\ln f(x, \alpha) = A_1(\alpha)\,\Psi(x) + A_0(\alpha) + \Psi_0(x).$$

Statistica exhaustivă este dată de relația

$$\alpha_1(\vec{x}) = \sum_{i=1}^{n} \Psi(x_i)$$

și evident de orice funcție de $\alpha_1(\vec{x})$.
Din ecuația verosimilității maxime, obținem

$$\sum_{i=1}^{n} \Psi(x_i) = -\frac{A_0'(\alpha^*)}{A_0'(\alpha)}\, n,$$

sau

$$\alpha^* = F\left(\sum_{i=1}^{n} \Psi(x_i)\right) = F(\alpha_1).$$

Prin urmare, dacă există o funcție de estimație suficientă α_1, orice soluție a ecuației de verosimilitate este funcție de α_1.

Aceste proprietăți ne arată importanța metodei verosimilității maxime. Metoda verosimilității maxime mai prezintă însă interes și prin faptul că ne poate da, în condiții destul de largi, o funcție de estimație asimptotic eficientă, a cărei definiție o dăm mai jos.

Să admitem că $\bar{\alpha}(x_1, x_2, \ldots, x_n)$ este o funcție de estimație corectă a lui α, astfel încît să avem

$$D(\bar{\alpha}) \simeq \frac{a}{\sqrt{n}},$$

unde a este o constantă.

Eficiența lui $\bar{\alpha}$ am definit-o prin relația

$$e(\bar{\alpha}) = \frac{1}{nD^2(\bar{\alpha}) \int_{-\infty}^{+\infty} \left[\frac{\partial \ln f(x, \alpha)}{\partial \alpha}\right]^2 f(x, \alpha)\, dx} = \frac{1}{nD^2(\bar{\alpha})\, M\left[\frac{\partial \ln f(x, \alpha)}{\partial \alpha}\right]^2}.$$

Rezultă că în cazul nostru avem

$$\lim_{n\to\infty} e(\bar{\alpha}) = \frac{1}{aM\left[\frac{\partial \ln f(x, \alpha)}{\partial \alpha}\right]^2}.$$

Această expresie poartă numele de *eficiență asimptotică* pentru $\bar{\alpha}$ și dacă ea este egală cu 1 vom spune că $\bar{\alpha}$ este o *funcție de estimație asimptotic eficientă*.

Fie

$$\alpha^* = \alpha^*(x_1, x_2, \ldots, x_n)$$

o soluție a ecuației verosimilității maxime.

Ne propunem să arătăm că, dacă

1. $\frac{1}{n}\left[\frac{\partial^2 \ln P}{\partial \alpha^2}\right]_{\alpha=\alpha^*}$ converge în probabilitate către o valoare finită A;

2. $\frac{1}{n}\left[\frac{\partial^3 \ln P}{\partial \alpha^3}\right]_{\alpha=\alpha^*}$ converge în probabilitate către o valoare finită B;

3. $\sqrt{n}(\alpha^* - \alpha)$ tinde la limită către o lege de probabilitate nedegenerată,

atunci variabila $\sqrt{n}\,(\alpha^*-\alpha)$ tinde, pentru $n\to\infty$, către o distribuție normală.

In relația

$$\frac{\partial \ln P(x,\ \alpha)}{\partial\alpha}=\sum_{i=1}^{n}\frac{\partial \ln f(x_i,\ \alpha)}{\partial\alpha},$$

să introducem notația

$$X_i=\frac{\partial \ln f(x_i,\ \alpha)}{\partial\alpha}.$$

Rezultă

$$\frac{\partial \ln P(x,\ \alpha)}{\partial\alpha}=\sum_{i=1}^{n}X_i.$$

Pe de altă parte,

$$M(X_i)=M\left(\frac{\partial \ln f(x_i,\ \alpha)}{\partial\alpha}\right)=\int_{-\infty}^{+\infty}\frac{\partial f(x,\ \alpha)}{\partial\alpha}\,dx=0.$$

Să presupunem că

$$D^2(X_i)=\int_{-\infty}^{+\infty}\left[\frac{\partial \ln f(x,\ \alpha)}{\partial\alpha}\right]^2 dx=h^2,$$

h fiind un număr finit, diferit de zero.

Atunci, după teorema limită centrală a calculului probabilităților, rezultă că $\sum_{i=1}^{n}X_i$ urmează la limită o distribuție normală.

Prin urmare

$$\lim_{n\to\infty}P\left[\frac{1}{h\sqrt{n}}\frac{\partial \ln P}{\partial\alpha}<\alpha\right]=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{a}e^{-\frac{x^2}{2}}\,dx. \qquad \text{(XVI.7,2)}$$

Dezvoltînd în serie după formula lui Taylor, avem

$$\frac{1}{h\sqrt{n}}\frac{\partial \ln P}{\partial\alpha}=\frac{1}{h\sqrt{n}}\left(\frac{\partial \ln P}{\partial\alpha}\right)_{\alpha^*}+\frac{1}{h\sqrt{n}}(\alpha-\alpha^*)\left(\frac{\partial^2 \ln P}{\partial\alpha^2}\right)_{\alpha^*}+$$
$$+\frac{1}{2h\sqrt{n}}(\alpha-\alpha^*)^2\left(\frac{\partial^3 \ln P}{\partial\alpha^3}\right)_{\alpha_0},$$

unde α_0 este un punct din intervalul $(\alpha,\ \alpha^*)$.

Avem însă, conform ipotezei,

$$\left(\frac{\partial \ln P}{\partial \alpha}\right)_{\alpha^*}=0,$$

astfel că vom putea scrie relația de mai sus sub forma

$$\frac{1}{h\sqrt{n}}\frac{\partial \ln P}{\partial \alpha}=[-h\sqrt{n}\,(\alpha-\alpha^*)]+\frac{1}{h}\left[h^2+\frac{1}{n}\left(\frac{\partial^2 \ln P}{\partial \alpha^2}\right)_{\alpha^*}\right]\sqrt{n}\,(\alpha-\alpha^*)+$$

$$+\frac{1}{2h\sqrt{n}}(\alpha-\alpha^*)^2\left(\frac{\partial^3 \ln P}{\partial \alpha^3}\right)_{\alpha^*}. \qquad \text{(XVI.7,3)}$$

Admițînd că primele două derivate în raport cu α ale funcției $f(x, \alpha)$ există, avem identitatea

$$f\frac{\partial^2 \ln f}{\partial \alpha^2}+f\left(\frac{\partial \ln f}{\partial \alpha}\right)^2=\frac{\partial^2 f}{\partial \alpha^2}.$$

Integrînd, rezultă

$$M\left[\frac{\partial^2 \ln f}{\partial \alpha^2}\right]+M\left(\frac{\partial \ln f}{\partial \alpha}\right)^2=\int_{-\infty}^{+\infty}\frac{\partial^2 f}{\partial \alpha^2}\,dx.$$

Insă, derivînd de două ori în raport cu α, egalitatea

$$\int_{-\infty}^{+\infty} f(x, \alpha)\,dx=1,$$

rezultă

$$\int_{-\infty}^{+\infty}\frac{\partial^2 f}{\partial \alpha^2}\,dx=0.$$

Prin urmare

$$M\left(\frac{\partial^2 \ln f}{\partial \alpha^2}\right)=-M\left(\frac{\partial \ln f}{\partial \alpha}\right)^2=-h^2,$$

$$M\left[h^2+\frac{1}{n}\left(\frac{\partial^2 \ln P}{\partial \alpha^2}\right)_{\alpha^*}\right]=0.$$

Scanned with OKEN Scanner

Să notăm

$$-h\sqrt{n}\,(\alpha-\alpha^*)=X,$$

$$\frac{1}{h}\left[h^2+\frac{1}{n}\left(\frac{\partial^2 \ln P}{\partial\alpha^2}\right)_{\alpha^*}\right]=Y,$$

$$\frac{1}{2h\sqrt{n}}(\alpha\cdot-\alpha^*)^2\left(\frac{\partial^3 \ln P}{\partial\alpha^3}\right)_{\alpha^*}=Z.$$

In baza proprietăţilor arătate pînă acum, rezultă că distribuţia variabilei (XVI.7,3) în spaţiul X, Y, Z tinde, pentru $n\to\infty$, către o distribuţie unidimensională şi că putem scrie

$$\lim_{n\to\infty} P\left[\frac{1}{h\sqrt{n}}\frac{\partial \ln P}{\partial\alpha}<a\right]=\lim_{n\to\infty} P\,(X+Y+Z<a)=\lim_{n\to\infty} P_n\,(X<a).$$

Ţinînd seama de (XVI.7,2), rezultă

$$\lim_{n\to\infty} P[h\sqrt{n}\,(\alpha^*-\alpha)<a]=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{a} e^{-\frac{x^2}{2}}\,dx. \qquad \text{(XVI.7,4)}$$

Teorema este deci demonstrată.
Din (XVI.7,4) rezultă

$$\lim_{n\to\infty} h^2 n D^2\,(\alpha^*)=1.$$

Avem însă

$$e\,(\alpha^*)=\frac{1}{nD^2(\alpha^*)\,M\left(\frac{\partial \ln f}{\partial\alpha}\right)^2}=\frac{1}{nh^2D^2\,(\alpha^*)},$$

astfel că

$$\lim_{n\to\infty} e\,(\alpha^*)=1.$$

Rezultă astfel că α^* este o funcţie de estimaţie asimptotic eficientă.

§ 8. Metoda momentelor

Vechea metodă de ajustare analitică a lui K. Pearson, bazată pe egalitatea momentelor empirice cu momentele teoretice, poate da rezultate şi la estimaţia parametrilor.

Egalînd astfel un moment oarecare de ordinul p cu momentul empiric corespunzător, obținem

$$\frac{\sum_{i=1}^{n} x_i^p}{n} = \int_{-\infty}^{+\infty} x^p f(x, \alpha)\, dx = M_p,$$

din care putem deduce pe α în funcție de $x_1, x_2, \ldots, x_n$.

Proprietățile stabilite la teoria selecției pentru mediile empirice dau o justificare folosirii acestei metode.

Intr-adevăr am demonstrat că

$$M\left(\frac{\sum_{i=1}^{n} x_i^p}{n}\right) = M_p,$$

$$D^2\left(\frac{\sum_{i=1}^{n} x_i^p}{n}\right) = \frac{M_{2p} - M_p^2}{n}.$$

Prin urmare, după cum am mai arătat la teoria selecției, $\frac{1}{n}\sum x_i^p$ este o funcție de estimație absolut corectă.

Să presupunem acum că în alcătuirea lui $f(x, \alpha_1, \alpha_2, \ldots, \alpha_k)$ intră unul sau mai mulți parametri necunoscuți. Cazul mai multor parametri va fi tratat în capitolul următor.

Anticipăm totuși, asupra acestui caz, pentru că la metoda momentelor el nu prezintă nici o dificultate.

Rezultă că M_p este o funcție $M_p(\alpha_1, \alpha_2, \ldots, \alpha_k)$ de acești parametri. Insemnăm

$$\overline{M}_p = \frac{\sum x_i^p}{n}.$$

Avînd de estimat k parametri, în baza metodei momentelor vom egala k momente empirice $\overline{M}_p$ cu momentele corespunzătoare teoretice, adică vom avea de rezolvat ecuațiile

$$M_{p_i}(\alpha_1, \alpha_2, \ldots, \alpha_k) = \overline{M}_{p_i} \qquad (i = 1, 2, \ldots, k). \qquad \text{(XVI.8,1)}$$

Numerele întregi p_i le alegem după voia noastră. Din aceste k ecuații obținem, prin rezolvarea lor, k funcții de estimație $\bar{\alpha}_i$ pentru parametrii α_i $(i = 1, 2, \ldots, k)$.

In baza teoremei de la § 4, rezultă că, în condiţii destul de largi, aceste funcţii de estimaţie sînt absolut corecte.

Noi am aplicat această metodă la ajustarea distribuţiilor empirice cu ajutorul curbelor Pearson, înainte de a studia teoria selecţiei. Ne putem întreba acum în ce caz funcţiile de estimaţie astfel găsite sînt eficiente sau asimptotic eficiente.

In distribuţiile Pearson există patru parametri necunoscuţi. Pentru ajustarea făcută la cap. XI, § 1 am aplicat metoda momentelor, înlocuind primele patru momente teoretice prin momentele corespunzătoare empirice.

Pentru ca ecuaţia de verosimilitate maximă să ne conducă la egalităţi de tipul (XVI.8,1) este necesar ca funcţia de verosimilitate maximă să fie de forma

$$P = e^{\beta+\beta_1 \sum_{i=1}^{n} x_i + \beta_2 \sum_{i=1}^{n} x_i^2 + \beta_3 \sum_{i=1}^{n} x_i^3 + \beta_4 \sum_{i=1}^{n} x_i^4},$$

unde β_1, β_2, β_3, β_4 sînt funcţii liniare date de cei patru parametri α_1, α_2, α_3, α_4, de unde rezultă că

$$f(x;\ \alpha_1,\ \alpha_2,\ \alpha_3,\ \alpha_4) = ke^{\beta_1 x+\beta_2 x^2+\beta_3 x^3+\beta_4 x^4}, \qquad \text{(XVI.8,2)}$$

k fiind determinat prin relaţia

$$k \int_{-\infty}^{+\infty} e^{\beta_1 x+\beta_2 x^2+\beta_3 x^3+\beta_4 x^4}\, dx = 1.$$

Aceasta este forma generală a distribuţiilor pentru care aplicarea metodei momentelor, pentru primele patru momente, conduce la funcţii de estimaţie asimptotic eficiente.

Distribuţia normală face parte din familia (XVI.8,1). După cum ştim, cînd avem un singur parametru de estimat, această distribuţie conduce uneori chiar la funcţii de estimaţie eficiente.

§ 9. Cazul discontinuu

Să considerăm o familie de distribuţii statistice discontinue, depinzînd de un parametru α definit în modul următor: distribuţia se referă la valorile sale discrete a_i, în număr finit sau infinit, care pot fi luate respectiv cu probabilităţile $p_i(\alpha)$.

Atunci

$$p_i(\alpha) \geqq 0, \quad \sum_{i=1}^{k} p_i(\alpha) = 1.$$

Presupunem că $p_i(\alpha)$ este o funcție continuă de α pentru orice i, admițînd o derivată, de asemenea funcție continuă de α.
Efectuînd n experiențe independente, am obținut valorile $x_1, x_2, \ldots, x_n$ care vor servi ca observație informativă pentru estimația lui α.
Fie

$$\bar{\alpha}(x_1, x_2, \ldots, x_n) = \bar{\alpha}(\vec{x})$$

o funcție de estimație a lui α, convergentă în medie pătratică către α.
Admitem că mulțimea punctelor a_i este finită.
In cazul cînd funcția de estimație $\bar{\alpha}(\vec{x})$ este numai corectă, adică

$$M(\alpha(\vec{x})) = \bar{\alpha} + \theta(\alpha),$$

se demonstrează, în același mod ca în cazul continuu, că există minimul

$$M(\bar{\alpha} - \alpha)^2 \geqq \frac{[1 + \theta'(\alpha)]^2}{n \sum_{i=1}^{n} \left(\frac{\partial \ln p_i(\alpha)}{\partial \alpha} \right)^2 p_i(\alpha)},$$

$\theta'(\alpha)$ fiind derivata funcției $\theta(\alpha)$.
Dacă $\bar{\alpha}$ este o funcție de estimație absolut corectă, rezultă că

$$D^2(\alpha) \geqq \frac{1}{n \sum_{i=1}^{n} \left(\frac{\partial \ln p_i(\alpha)}{\partial \alpha} \right)^2 p_i(\alpha)}.$$

Noțiunea de eficiență se definește în același mod ca și pentru distribuțiile continue.
In cazul cînd mulțimea a_i este infinită, pentru ca formele de mai sus să fie valabile trebuie să fie îndeplinite și alte condiții.
Este de observat că în cazul discontinuu finit se poate construi ușor o funcție de estimație.

Scanned with OKEN Scanner

Referindu-ne la cazul cînd variabilele $x_i\,(i=1, 2, \ldots, n)$ pot lua numai k valori $a_i\,(i=1, 2, \ldots, k)$, cu probabilităţile respective $p_i(\alpha)$, $\alpha=1, 2, \ldots, k$, vom trata problema vectorial.

Fie $y\,(y_1, y_2, \ldots, y_k)$ variabila vectorială care ia valorile $e_1\,(1, 0, \ldots, 0)$, $e_2\,(0, 1, \ldots, 0), \ldots,$ $e_k\,(0, 0, \ldots, 0, 1)$ cu probabilităţile respective $p_1(\alpha), p_2(\alpha), \ldots, p_n(\alpha)$.

Introducem funcţia $\varphi(y; \alpha)$, care se bucură de următoarele proprietăţi

1. $M[\varphi]=0$;

2. $D^2[\varphi]=h^2$, unde h este un număr finit, diferit de zero, depinzînd de α;

3. $M\left[\frac{\partial\varphi}{\partial\alpha}\right]$ există şi nu se anulează într-un interval finit I care conţine pe φ_0, adevărata valoare, necunoscută, a lui φ.

In cele n experienţe efectuate, să presupunem că y a luat de n_1 ori valoarea e_1, de n_2 ori valoarea $e_2, \ldots,$ de n_k ori valoarea e_k $\left(\sum_{i=1}^{k} n_i=n\right)$.

Notăm

$$\overline{\varphi}(\alpha)=\frac{1}{n}\sum n_i\varphi(e_i, \alpha). \qquad \text{(XVI.9,1)}$$

Fie $\bar{\alpha}$ o rădăcină a ecuaţiei (XVI.9,1) în intervalul I.

Vom arăta că $\bar{\alpha}$ este o funcţie de estimaţie corectă a lui α.

Intr-adevăr, din (XVI.6,4) rezultă imediat că $\overline{\varphi}(\alpha)$ tinde în probabilitate către $M(\varphi)$ care, în baza lui 1, este zero.

Avem deci egalităţile

$$\overline{\varphi}(\bar{\alpha})=0, \quad M[\overline{\varphi}(\alpha_0)]=0.$$

Repetînd atunci un raţionament asemănător cu acel făcut în § 7 pentru a demonstra că soluţia ecuaţiei verosimilităţii maxime tinde în probabilitate către adevărata valoare a parametrului α, vom obţine

$$P\left[\frac{\varphi(\alpha_0)}{M\left[\frac{\partial\varphi}{\partial\alpha}\right]_{\alpha_0}}<a\right]=P(\alpha_0-\bar{\alpha}<a). \qquad \text{(XVI.9,2)}$$

Deoarece $M[\varphi(\alpha_0)] = 0$, rezultă mai departe

$$\frac{\varphi(\alpha_0)}{M\left[\frac{\partial\varphi}{\partial\alpha}\right]_{\alpha_0}} \doteq \frac{\varphi(\alpha_0) - M[\varphi(\alpha_0)]}{M\left[\frac{\partial\varphi}{\partial\alpha}\right]_{\alpha_0}} = \frac{\sum\left[\frac{n_i}{n} - p_i(\alpha_0)\right]\varphi(e_i, \alpha_0)}{M\left[\frac{\partial\varphi}{\partial\alpha}\right]_{\alpha_0}}. \quad \text{(XVI.9,3)}$$

In baza teoremei lui Liapunov, rezultă atunci că variabila (XVI.9,3) tinde, pentru $n \to \infty$, către o distribuție normală.
Avem deci, ținînd seamă de (XVI.9,2) și (XVI.9,3),

$$M[\alpha] = \alpha_0,$$

$$D(\bar{\alpha}) = \frac{D(\varphi_0)}{\sqrt{n}\, M\left[\frac{\partial\varphi}{\partial\alpha}\right]};$$

$\bar{\alpha}$ este deci o funcție de estimație corectă a lui α și, mai mult decît atît, din considerațiile de mai sus, reiese că este și laplaciană.
Să considerăm funcția particulară

$$\Psi(y; \alpha) = \sum_{i=1}^{k} \frac{p_i'(\alpha)}{p_i(\alpha)} y_i.$$

Să vedem în ce condiții ea poate satisface proprietățile 1, 2, 3.
Avem

$$M[\Psi(y; \alpha)] = \sum_{i=1}^{k} \frac{p_i'(\alpha)}{p_i(\alpha)} p_i(\alpha) = \sum_{i=1}^{k} p_i'(\alpha) = 0,$$

deoarece

$$\sum_{i=1}^{k} p_i(\alpha) = 1$$

și prin derivare se obține

$$\sum_{i=1}^{k} p_i'(\alpha) = 0, \quad \sum_{i=1}^{k} p_i''(\alpha) = 0.$$

Condiția 1 este deci îndeplinită.
Pe de altă parte

$$D^2[\Psi] = \sum_{i=1}^{k} \left[\frac{p_i'(\alpha)}{p_i(\alpha)}\right]^2 p_i(\alpha) = \sum_{i=1}^{k} \frac{[p_i'(\alpha)]^2}{p_i(\alpha)}.$$

Scanned with OKEN Scanner

Pentru ca să fie îndeplinită condiţia 2, trebuie ca expresia de mai sus să fie finită şi nenulă în intervalul I

$$M\left[\frac{\partial \Psi}{\partial \alpha}\right] = \sum_{i=1}^{k}\left[\frac{p_i''(\alpha)}{p_i(\alpha)} - \frac{[p_i'(\alpha)]^2}{p_i^2(\alpha)}\right] p_i(\alpha) =$$

$$= \sum_{i=1}^{k} p_i''(\alpha) - \sum_{i=1}^{k} \frac{[p_i'(\alpha)]^2}{p_i(\alpha)} = -D^2(\Psi) \neq 0.$$

Deci condiţia 3 rezultă din 2.

Rezultă atunci că rădăcina ecuaţiei

$$\frac{1}{n}\sum n_i \frac{p_i'(\alpha)}{p_i(\alpha)} = 0 \qquad \text{(XVI.9,4)}$$

ne dă o funcţie de estimaţie corectă.

Determinarea unei funcţii de estimaţie prin rezolvarea ecuaţiei (XVI.9,1) reprezintă în realitate o extindere a principiului verosimilităţii maxime. Cazul particular al ecuaţiei (XVI.9,1), dat de ecuaţia (XVI.9,2), rezultă chiar din aplicarea metodei verosimilităţii maxime.

Intr-adevăr, probabilitatea ca să obţinem în n probe de n_i ori vectorul e_i $(i = 1, 2, \ldots, k)$ este

$$P(\alpha) = \frac{n!}{n_1!\, n_2! \ldots n_k!} p_1^{n_1}(\alpha) \cdot p_2^{n_2}(\alpha) \ldots, p_k^{n_k}(\alpha) \quad (n_1 + n_2 + \ldots + n_k = n).$$

Ecuaţia verosimilităţii maxime

$$\frac{\partial \ln p(\alpha)}{\partial \alpha} = 0$$

conduce în acest caz la (XVI.9,4).

Rezultă atunci din teoria generală că funcţia de estimaţie dată de (XVI.9,4) este asimptotic eficientă.

Eficacitatea funcţiilor de estimaţie obţinute din ecuaţia (XVI.9,1) este cea mai mare, din punct de vedere asimptotic, cînd se foloseşte funcţia de estimaţie dedusă din (XVI.9,4).

Adeseori se foloseşte pentru estimaţie o metodă bazată pe expresia lui χ^2. Vom arăta cum se aplică această metodă în cazul discontinuu finit.

Scanned with OKEN Scanner

Avem în cazul de față

$$\chi^2 = \sum_{i=1}^{k} \frac{[n_i - np_i(\alpha)]^2}{np_i(\alpha)}.$$

Vom estima parametrul α punînd condiția ca χ^2 să fie minim. Obținem astfel o ecuație

$$2\sum_{i=1}^{k} \frac{\left[\frac{n_i}{n} - p_i(\alpha)\right] p_i'(\alpha)}{p_i'(\alpha)} + \sum_{i=1}^{k} \frac{\left[\frac{n_i}{n} - p_i(\alpha)\right]^2 p_i'(\alpha)}{p_i^2(\alpha)} = 0. \qquad \text{(XVI.9,5)}$$

Ecuația (XVI.9,5) este în general greu de rezolvat. Din punct de vedere asimptotic, ea conduce însă la același rezultat ca și ecuația verosimilității maxime, după cum se poate vedea ușor.

Capitolul XVII

Teoria estimației pentru familiile de distribuții cu mai mulți parametri

§ 1. Considerații generale. Definiții

Examinăm, în acest capitol, problema estimației în cazul familiilor de distribuții statistice dintr-un spațiu m-dimensional R_m, care depind de un număr oarecare de parametri. Pentru aceasta, întrebuințăm limbajul folosit de E. B. Dînkin, care permite o expunere sistematică a rezultatelor obținute pînă acum în acest domeniu.

Fie S un domeniu al spațiului numeric R_s. Definim o familie de distribuții statistice în spațiul m-dimensional R_m, considerînd, pentru fiecare $\theta \in S$ o distribuție din R_m.

O asemenea familie $f(\bar{x}, \theta)$ este regulată într-un domeniu G din R_m dacă, pentru fiecare $\theta \in S$, $f(\bar{x}, \theta)$ admite o densitate $p(\vec{x}, \theta)$, cînd $\bar{x} \in G$. Presupunem că în întreg domeniul $G \times S$ derivatele parțiale $\frac{\partial p(\vec{x}, \theta)}{\partial x_i}$ există și sînt continue.

Funcția $K(\vec{x})$, definită în G și luînd valori într-o mulțime T, constituie o statistică suficientă în G pentru familia de distribuții considerată, dacă

$$p(\vec{x}, \theta) = \overline{p}(K(\vec{x}), \theta) q(\vec{x}) \qquad (\vec{x} \in G, \ \theta \in S).$$

Fie $K_1(\vec{x})$ și $K_2(\vec{x})$ două statistici suficiente. Vom spune că statistica $K_1(\vec{x})$ este *subordonată* statisticii $K_2(\vec{x})$ dacă din $K_2(\vec{x'}) = K_2(\vec{x''})$ rezultă $K_1(\vec{x'}) = K_1(\vec{x''})$.

Statisticile $K_1(\vec{x})$ și $K_2(\vec{x})$ sînt *echivalente* dacă fiecare dintre ele este subordonată celeilalte.

Statistica $\varepsilon(\vec{x})$ este suficientă pentru orice familie de distribuții.

Vom spune că o statistică $K(\vec{x})$ este *banală* în G dacă este echivalentă cu $\varepsilon(\vec{x})$.

Funcția $K(\vec{x})$ este o statistică *necesară* dacă este subordonată oricărei statistici suficiente.

Din definiție rezultă că orice funcție subordonată unei statistici necesare este de asemenea o statistică necesară. De asemenea, rezultă că două statistici necesare și suficiente în același timp sînt echivalente.

Problema estimației se pune în același mod ca și pentru un singur parametru. Se fac n observații independente și pe baza lor se găsește o valoare apropiată numeric pentru parametrii necunoscuți. Aceste n observații independente ne-au condus la punctele $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n$ respectiv din R_m. Rezultatele observațiilor formează un spațiu $R_{mn} = R_m \times R_m \times \ldots \times R_m$, ale cărui elemente sînt sistemele ordonate de puncte $(\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n)$ ale spațiului R_m.

Deoarece $p(\vec{x}, \theta)$ este o densitate de repartiție pentru $\vec{x} \in G$, densitatea de repartiție pentru punctul $(x_1, x_2, \ldots, x_n)$ din R_{mn} este $p(\vec{x}_1, \theta), p(\vec{x}_2, \theta), \ldots, p(\vec{x}_n, \theta)$, definind astfel o familie de distribuții statistice în acest spațiu.

Funcțiile $K(\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n)$ definite în domeniul

$$\vec{x}_1 \in G, \vec{x}_2 \in G, \ldots, \vec{x}_n \in G$$

sînt statistici necesare și suficiente în acest domeniu, în virtutea aceleiași definiții ca și pentru familia de distribuții statistice definite în G.

§ 2. Determinarea unei statistici necesare și suficiente

Se poate găsi ușor o statistică necesară și suficientă în G. Pentru aceasta, să alegem un $\theta \in S$ și să considerăm funcția

$$g_{\vec{x}}(\theta) = \ln p(\vec{x}, \theta) - \ln p(\vec{x}, \theta_0).$$

Funcția $g_x(\theta)$ este o statistică necesară și suficientă.

Relația precedentă se poate scrie

$$p(\vec{x}, \theta) = p(\vec{x}, \theta_0)\, e^{g_{\vec{x}}(\theta)}.$$

Prin urmare, $g_{\vec{x}}(\theta)$ este o statistică suficientă.

Pe de altă parte, dacă $K(\vec{x})$ este o statistică suficientă, avem, prin definiție,

$$p(\vec{x}, \theta) = \overline{p}(K(\vec{x}), \theta)\, q(\vec{x}),$$

$$p(\vec{x}, \theta_0) = \overline{p}(K(\vec{x}), \theta_0)\, q(\vec{x}),$$

$$g_{\vec{x}}(\theta) = \ln \overline{p}(K(\vec{x}), \theta) - \ln \overline{p}(K(\vec{x}), \theta_0).$$

Deci $g_{\vec{x}}(\theta)$ este o statistică subordonată lui $K(\vec{x})$. Propoziția este deci demonstrată.

Scanned with OKEN Scanner

Densitatea punctului $(\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n)$ din R_{mn} fiind

$$\prod_{i=1}^{n} p(\vec{x}_i, \theta),$$

rezultă că

$$g_{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n}(\theta) = g_{\vec{x}_1}(\theta) + g_{\vec{x}_2}(\theta) + \ldots + g_{\vec{x}_n}(\theta)$$

este o statistică necesară şi suficientă pentru grupul de informaţii $(\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n)$ din R_{mn}.

Folosind terminologia din paragraful precedent, se vede că $g_{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n}(\theta)$ nu reprezintă altceva decît o statistică exhaustivă pentru grupul de informaţii considerat.

§ 3. Cazul familiilor de distribuţii unidimensionale

In cazul cînd familia de distribuţii este definită pe R, teoremele pe care le-am expus mai sus se pot generaliza. Se rezolvă astfel problema estimaţiei a mai mulţi parametri necunoscuţi dintr-o lege specială.

Admitem, ca şi pînă acum, că densitatea de distribuţie $p(x, \theta)$ este o funcţie continuă şi pozitivă într-un interval Δ şi admite o derivată continuă pentru orice $\theta \in S$. Familia de distribuţii statistice este definită pe Δ.

Să notăm atunci cu $L(\Delta)$ spaţiul liniar minimal al funcţiilor definite pe Δ, conţinînd constantele şi funcţiile $g_x(\theta)$ pentru orice $\theta \in S$. Fie $r+1$ dimensiunea lui $L(\Delta)$.

Să considerăm cazul cînd r este finit. In acest caz, pot fi demonstrate următoarele propoziţii:

O statistică suficientă referitoare la un grup de n informaţii este banală pentru orice $n \leq r$.

Dacă funcţiile $1, \varphi_1(x), \ldots, \varphi_r(x)$ *formează o bază în* $L(\Delta)$, *atunci sistemul de funcţii*

$$K_i(x_1, x_2, \ldots, x_n) = \varphi_i(x_1) + \varphi_i(x_2) + \ldots + \varphi_i(x_n) \qquad (i = 1, 2, \ldots, r)$$

constituie o statistică necesară şi suficientă referitoare la grupul de n informaţii pentru orice $n \geq r$, *funcţiile* K_i *fiind liniar independente.*

Să considerăm mai întîi propoziția a doua. Deoarece $\varphi(x) \in L$, $g_x(\theta) \in L$, pentru orice $\theta_q \in S$, atunci

$$\varphi(x) = \sum_{q=1}^{l} c_q g_x(\theta_q) + c_0,$$

$$K(x_1, x_2, \ldots, x_n) = \sum_{q=1}^{l} c_q g_x(\theta_q) + nc_0.$$

Prin urmare $K_i(x_1, x_2, \ldots, x_n)$ este subordonată statisticii necesare $g_{x_1, x_2, \ldots, x_n}(\theta)$.

Pe de altă parte,

$$g_x(\theta) = \sum_{j=1}^{s} a_j(\theta)\, \varphi_j(x) + a_0(\theta),$$

$$g_{x_1, x_2, \ldots, x_n}(\theta) = \sum_{j=1}^{s} a_j(\theta)\, K_j(x_1, x_2, \ldots, x_n) + na_0(\theta).$$

Sistemul $(K_1, K_2, \ldots, K_n)$ formează o statistică suficientă, deoarece subordonează statistica suficientă $g_{x_1, x_2, \ldots, x_n}(\theta)$.

Am demonstrat că, pentru $n \geqslant r$, K reprezintă o statistică necesară și suficientă. Mai rămîne să arătăm că funcțiile K_i sînt liniar independente. Avem, evident,

$$\frac{D(K_1, K_2, \ldots, K_s)}{D(x_1, x_2, \ldots, x_s)} = \begin{vmatrix} \varphi_1'(x_1) & \varphi_1'(x_2) \ldots \varphi_1'(x_s) \\ \varphi_2'(x_1) & \varphi_2'(x_2) \ldots \varphi_2'(x_s) \\ \ldots & \ldots \ldots \ldots \\ \varphi_s'(x_1) & \varphi_s'(x_2) \ldots \varphi_s'(x_s) \end{vmatrix},$$

funcțiile $\varphi_i(x_j)$ fiind continue în Δ.

Din egalitatea

$$\frac{D(K_1, K_2, \ldots, K_s)}{D(x_1, x_2, \ldots, x_s)} = 0$$

rezultă dependența liniară a sistemului $1, \varphi_1, \varphi_2, \ldots, \varphi_s$, ceea ce este contrar ipotezei asupra alegerii funcțiilor φ_i.

Rămîne să dovedim prima propoziție.

Dacă $n \leqslant r$, putem găsi în L funcțiile $\varphi_1, \varphi_2, \ldots, \varphi_n$ astfel încît sistemul $1, \varphi_1, \varphi_2, \ldots, \varphi_n$ să fie liniar independent. În virtutea primei propoziții, vom găsi corespunzător acestui sistem n funcții $K_i(x)$ pentru care determinantul funcțional

$$\frac{D(K_1, K_2, \ldots, K_n)}{D(x_1, x_2, \ldots, x_n)}$$

nu este identic nul.

Atunci

$$x_i = x_i(K_1, K_2, \ldots, K_n) \qquad (i = 1, 2, \ldots, n)$$

în vecinătatea unui punct, unde determinantul de mai sus este diferit de zero. Statistica banală este deci subordonată oricărei statistici necesare, ceea ce demonstrează propoziția.

§ 4. Determinarea densității de distribuție în cazul unei statistici necesare și suficiente de rang finit

Vom numi *rang* al unei familii de distribuții statistice cel mai mare număr r pentru care nu avem decît statistici suficiente banale cînd $n \leqslant r$.

În acest caz, condiția necesară și suficientă pentru ca să existe o statistică necesară și suficientă de rangul r este ca să avem relația

$$\ln p(x, \theta) = \sum_{i=1}^{r} \varphi_i(x) c_i(\theta) + c_0(\theta) + \varphi_0(x) \qquad (x \in \Delta, \ \theta \in S).$$

Funcțiile $(1, \varphi_1, \varphi_2, \ldots, \varphi_r)$ și $(1, c_1, c_2, \ldots, c_r)$ sînt liniar independente. Se vede că relația de mai sus extinde formula pe care am stabilit-o în cazul familiilor de distribuții cu un singur parametru și la familiile de distribuții cu mai mulți parametri.

Ordinul minim al statisticii exhaustive coincide cu rangul statisticii necesare și suficiente.

Condiția este necesară. Fie $1, \varphi_1, \varphi_2, \ldots, \varphi_r$ o bază în $L(\Delta)$. Atunci

$$g_x(\theta) = \sum_{i=1}^{r} c_i(\theta) \varphi_i(x) + c_0(\theta) \qquad (\theta \in S).$$

Scanned with OKEN Scanner

Pe de altă parte,

$$g_x(\theta)=\ln p(x,\ \theta)-\ln p(x,\ \theta_0).$$

Din aceste două relații rezultă imediat valoarea lui $\ln p(x, \theta)$ sub forma dată mai sus.

Mai trebuie să demonstrăm că sistemul de funcții $1, c_1, c_2, \ldots, c_r$ este liniar independent. Intr-adevăr, în caz contrar

$$c_r(\theta)=b_0+b_1c_1(\theta)+\ldots+b_{r-1}c_{r-1}(\theta)$$

și

$$g_x(\theta)=\sum_{i=1}^{r-1}c_i(\theta)\,\Psi_i(x)+c_0(\theta)+b_0\varphi_r(x),$$

unde

$$\Psi_i(x)=\varphi_i(x)+b_i\varphi_r(x) \qquad (i=1,\ 2,\ldots,r-1).$$

Insă

$$g_x(\theta)=\sum_{i=1}^{r-1}c_i(\theta)\,\Psi_i(x)+c_0(\theta_0)+b_0\,\varphi_r(x),$$

$$g_x(\theta_0)=0,$$

după cum rezultă din expresia lui $g_x(\theta)$ în raport cu $p(x, \theta)$. Deci

$$g_x(\theta)=p(x,\ \theta)-p(x,\ \theta_0)=$$

$$=\sum_{i=1}^{r-1}[c_i(\theta)-c_i(\theta_0)]\,\Psi_i(x)+c_0(\theta)-c_0(\theta_0).$$

Această relație este însă imposibilă, deoarece numărul dimensiunilor lui L nu poate fi mai mic decît $r+1$, iar $1, \varphi_1, \varphi_2, \ldots, \varphi_r$ formează o bază în acest spațiu liniar. Condiția este suficientă. Intr-adevăr

$$g_x(\theta)=\ln p(x,\ \theta)-\ln p(x,\ \theta_0),$$

$$g_x(\theta)=\sum_{i=1}^{r}[c_i(\theta)-c_i]\,\varphi_i(x)+c_0(\theta)-c_0,$$

unde $c_i=c_i(\theta_0)$. Se vede de aici că numărul dimensiunilor lui L nu este mai mare decît $r+1$.

Funcțiile $1, c_1(\theta), \ldots, c_r(\theta)$ fiind liniar independente, rezultă că și funcțiile $c_1(\theta)-c_1,\ c_2(\theta)-c_2, \ldots, c_r(\theta)-c_r$ sînt liniar independente.

Putem alege atunci $\theta_1 \in S$, $\theta_2 \in S, \ldots, \theta_r \in S$ astfel încît $\| c_i(\theta_j) - c_i \|$ să fie diferită de zero.

Din relațiile de mai sus rezultă atunci că

$$\varphi_i(x) = \sum_{j=1}^{r} b_i^j g_x(\theta_j) + b_i^0,$$

b_i^j fiind niște constante și

$$K_i(x_1, x_2, \ldots, x_n) = \sum_{j=1}^{r} b_i^j g_{x_1, x_2, \ldots, x_n}(\theta_j) + b_i^0.$$

Din relațiile de mai sus rezultă că funcțiile 1, $\varphi_1, \varphi_2, \ldots, \varphi_r$, fiind liniar independente, formează o bază în L. In consecință, după cum am arătat în prima parte, K_i reprezintă o statistică necesară și suficientă referitoare la grupul considerat.

§ 5. Metoda verosimilității maxime

Să presupunem că sîntem în cazul unei distribuții unidimensionale, însă există un număr finit de parametri $\alpha_1, \alpha_2, \ldots, \alpha_k$.

Notăm prin $f(x; \alpha_1, \alpha_2, \ldots, \alpha_k)$ densitatea distribuției statistice.

Efectuînd un număr de n experiențe independente, am obținut valorile de selecție

$$x_1, \quad x_2, \ldots, x_n.$$

Considerăm funcția de verosimilitate

$$P = \prod_{i=1}^{n} f(x_i; \alpha_1, \alpha_2, \ldots, \alpha_k) \qquad \text{(XVII.5,1)}$$

și ecuațiile de verosimilitate

$$\frac{\partial \ln P}{\partial \alpha_i} = 0 \qquad (i = 1, 2, \ldots, k). \qquad \text{(XVII.5,2)}$$

Prin rezolvarea sistemului (XVII.5,2) vom găsi un sistem de soluții $\alpha_1^*, \alpha_2^*, \ldots, \alpha_k^*$ care reprezintă, în condiții similare cu cele date pentru cazul unui singur parametru, funcții de estimație corecte pentru parametrii $\alpha_1, \alpha_2, \ldots, \alpha_k$.

Scanned with OKEN Scanner

Să notăm prin $\alpha_1^0, \alpha_2^0, \ldots, \alpha_k^0$ adevăratele valori ale parametrilor $\alpha_1, \alpha_2, \ldots, \alpha_k$ și prin

$$\left[\frac{\partial \ln f}{\partial \alpha_k}\right]_0$$

valoarea derivatei $\frac{\partial \ln f(x; \alpha_1, \alpha_2, \ldots, \alpha_k)}{\partial \alpha_k}$ în punctul $\alpha_1^0, \alpha_2^0, \ldots, \alpha_k^0$.

Atunci, presupunînd că $f(x; \alpha_1, \alpha_2, \ldots, \alpha_k)$ este continuă în x, că într-un interval deschis k-dimensional, I, conținînd punctul $\alpha_1^0, \alpha_2^0, \ldots, \alpha_k^0$, derivata $\frac{\partial f}{\partial \alpha_k}$ este continuă în α_k pentru orice x, că ea tinde, cînd $n \to \infty$, către o funcție continuă de α_k și că $\frac{\partial f}{\partial \alpha_k}$ nu se anulează în I, rezultă că variabilele

$$\sqrt{n}\,(\alpha_1^* - \alpha_1^0), \quad \sqrt{n}\,(\alpha_2^* - \alpha_2^0), \ldots, \quad \sqrt{n}\,(\alpha_k^* - \alpha_k^0)$$

tind, pentru $n \to \infty$, către o distribuție normală, cu centrul în origine și matricea momentelor fiind egală cu A^{-1}, unde am însemnat prin $A = \| a_{ij} \|$ matricea formată cu elementele

$$a_{ij} = \int_{-\infty}^{+\infty} \left(\frac{\partial \ln f}{\partial \alpha_i}\right)_0 \left(\frac{\partial \ln f}{\partial \alpha_j}\right)_0 dx.$$

Nu mai este nevoie să prezentăm demonstrația acestei proprietăți deoarece nu diferă esențial de aceea pe care am dat-o în cazul unui singur parametru.

§ 6. Funcții de estimație eficiente

Am văzut în cazul unui singur parametru că, în anumite condiții, există o funcție de estimație corectă pentru care dispersia atinge un minim. Această funcție de estimație a fost numită funcție de estimație eficientă.

Problema se pune în același mod și pentru cazul mai multor parametri.

Fie $f(x; \alpha_1, \alpha_2, \ldots, \alpha_k)$ densitatea unei distribuții unidimensionale cu k parametri necunoscuți $\alpha_1, \alpha_2, \ldots, \alpha_k$. Pentru tratarea acestei

chestiuni, numai notaţia diferă de la cazul distribuţiilor cu mai multe dimensiuni la o singură dimensiune. De aceea pentru simplificarea scrierii ne limităm la o singură dimensiune.

Efectuînd o selecţie de n probe independente am găsit valorile

$$x_1,\ x_2, \ldots, x_n.$$

Notăm prin P funcţia de verosimilitate

$$P(x_1, \ldots, x_n;\ \alpha_1,\ \alpha_2, \ldots, \alpha_k) = \prod_{i=1}^{n} f(x_i;\ \alpha_1,\ \alpha_2, \ldots, \alpha_k).$$

Admitem că parametrii $\alpha_1, \alpha_2, \ldots, \alpha_k$ variază într-un interval k-dimensional deschis I din spaţiul parametrilor S.

Este evident că avem

$$\int_{-\infty}^{+\infty} \ldots \int_{-\infty}^{+\infty} P dx_1 dx_2 \ldots dx_n = 1. \qquad \text{(XVII.6,1)}$$

De aici rezultă prin derivare în raport cu parametrul α_i

$$\int_{-\infty}^{+\infty} \ldots \int_{-\infty}^{+\infty} \frac{\partial P}{\partial \alpha_i}\, dx_1 dx_2 \ldots dx_n = 0, \qquad (i = 1,\ 2, \ldots, k),$$

admiţînd că această integrală există.

Se vede imediat că relaţia de mai sus poate fi scrisă

$$M\left[\frac{\partial \ln P}{\partial \alpha_i}\right] = 0 \qquad (i = 1,\ 2, \ldots, k). \qquad \text{(XVII.6,2)}$$

Fie

$$\bar{\alpha}_1 = \bar{\alpha}_1(x_1,\ x_2, \ldots, x_n), \quad \bar{\alpha}_2 = \bar{\alpha}_2(x_1, \ldots, x_n), \ldots, \bar{\alpha}_k = \bar{\alpha}_k(x_1, \ldots, x_n)$$

un sistem de k funcţii de estimaţie corectă a parametrilor $\alpha_1, \alpha_2, \ldots, \alpha_k$.

Avem, prin urmare,

$$M(\bar{\alpha}_i) = \alpha_i \qquad (i = 1,\ 2, \ldots, k)$$

sau

$$\int_{-\infty}^{+\infty} \ldots \int_{-\infty}^{+\infty} \bar{\alpha}_i P dx_1 dx_2 \ldots dx_n = \alpha_i.$$

Dacă derivăm în raport cu α_j, obținem

$$\int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} \bar{\alpha}_i \frac{\partial P}{\partial \alpha_j} dx_1 dx_2 \dots dx_n = 0, \quad \text{pentru} \quad i \neq j,$$

$$\int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} \bar{\alpha}_i \frac{\partial P}{\partial \alpha_i} dx_1 dx_2 \dots dx_n = 1.$$

Aceste relații se pot scrie

$$M\left[\bar{\alpha}_i \frac{\partial \ln P}{\partial \alpha_j}\right] = 0, \quad \text{pentru} \quad i \neq j,$$

$$M\left[\bar{\alpha}_i \frac{\partial \ln P}{\partial \alpha_i}\right] = 1. \qquad \text{(XVII.6,3)}$$

Ținînd seamă de (XVII.6,2) obținem astfel

$$M\left[(\bar{\alpha}_i - \alpha_i) \frac{\partial \ln P}{\partial \alpha_i}\right] = 1 \qquad (i = 1, 2, \dots, k),$$

$$M\left[(\bar{\alpha}_i - \alpha_i) \frac{\partial \ln P}{\partial \alpha_j}\right] = 0 \quad \text{pentru} \quad j \neq i.$$

Să notăm

$$M(\bar{\alpha}_i - \alpha_i)(\bar{\alpha}_j - \alpha_j) = a_{ij},$$

$$M\left(\frac{\partial \ln P}{\partial \alpha_i} \cdot \frac{\partial \ln P}{\partial \alpha_j}\right) = n b_{ij}$$

și să admitem existența valorilor a_{ij} și b_{ij}.
Matricele $A = \| a_{ij} \|$ și $B = n \| b_{ij} \|$ sînt matrice de momente. Ele sînt deci pozitiv definite. Mai presupunem că rangul matricei A este k și că deci există inversa A^{-1}.
Să considerăm acum matricea momentelor pentru variabilele

$$\bar{\alpha}_1, \bar{\alpha}_2, \dots, \bar{\alpha}_k, \quad \frac{\partial \ln P}{\partial \alpha_1}, \quad \frac{\partial \ln P}{\partial \alpha_2}, \dots, \frac{\partial \ln P}{\partial \alpha_k}.$$

Ținînd seama de relațiile (XVII.6,3) și de notațiile noastre, rezultă că matricea momentelor, C, a variabilelor de mai sus este

$$C=\begin{Vmatrix} a_{11}\dots a_{1k} & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ a_{k1}\dots a_{kk} & 0 & \dots & 1 \\ 1 \dots 0 & nb_{11}\dots nb_{1k} & & \\ \dots & \dots & \dots & \dots \\ 0 \dots 1 & nb_{k1}\dots nb_{kk} & & \end{Vmatrix}.$$

Aplicînd regula de înmulțire a matricelor, avem

$$\begin{Vmatrix} a_{11}\dots a_{1k} & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ a_{k1}\dots a_{kk} & 0 & \dots & 1 \\ 1 \dots 0 & nb_{11}\dots nb_{1k} & & \\ \dots & \dots & \dots & \dots \\ 0 \dots 1 & nb_{k1}\dots nb_{kk} & & \end{Vmatrix} \cdot \begin{Vmatrix} a_{11}^{-1}\dots a_{1k}^{-1} & 1\dots 0 \\ \dots & \dots \\ a_{k1}^{-1}\dots a_{kk}^{-1} & 0\dots 1 \\ 0 \dots 0 & 1\dots 0 \\ \dots & \dots \\ 0 \dots 0 & 0\dots 1 \end{Vmatrix} =$$

$$= \begin{Vmatrix} 1 & 0 \dots 0 & 1 & 0 \dots 0 \\ 0 & 1 \dots 0 & 0 & 1 \dots 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 \dots 1 & 0 & 0 \dots 1 \\ a_{11}^{-1} & a_{12}^{-1}\dots a_{1k}^{-1} & nb_{11} & nb_{12}\dots nb_{1k} \\ \dots & \dots & \dots & \dots \\ a_{k1}^{-1} & a_{k2}^{-1}\dots a_{kk}^{-1} & nb_{k1} & nb_{k2}\dots nb_{kk} \end{Vmatrix}.$$

Să notăm matricea din membrul al doilea al acestei egalități prin D. Se observă, scăzînd respectiv primele k coloane din următoarele k coloane, că avem

$$|D|=|B-A^{-1}|.$$

Insă din produsul matriceal de mai sus, rezultă că

$$|D|=|C|\cdot|A^{-1}|.$$

De aici rezultă imediat, ținînd seamă că A și C sînt matrice pozitiv definite că

$$|D|\geqq 0$$

sau

$$|B-A^{-1}|\geqq 0.$$

Proprietatea de mai sus se păstrează și pentru minorii principali ai determinantului $|B-A^{-1}|$, după cum se poate vedea ușor repetînd același raționament pentru sistemele de variabile

$$\bar{\alpha}_1,\ \bar{\alpha}_2,\ldots,\bar{\alpha}_j,\quad \frac{\partial \ln P}{\partial \alpha_1},\ldots,\frac{\partial \ln P}{\partial \alpha_j}\qquad (i=1,\ 2,\ldots,k-1).$$

Prin urmare matricea $\|B-A^{-1}\|$ este definită pozitivă.
Dacă y este o matrice coloană oarecare

$$y=\begin{Vmatrix} y_1\\ y_2\\ \cdot\\ \cdot\\ \cdot\\ y_k\end{Vmatrix}$$

rezultă că forma pătratică $y'\,\|B-A^{-1}\|\,y$ este definită pozitivă. Avem prin urmare

$$y'\,\|B-A^{-1}\|\,y\geqq 0$$

sau

$$y'By'\geqq y'A^{-1}y.$$

Insă

$$y'A^{-1}y=k+2 \qquad \text{(XVII.6,4)}$$

reprezintă ecuația elipsoidului de concentrație al variabilelor

$$\bar{\alpha}_1,\ \bar{\alpha}_2,\ldots,\bar{\alpha}_k.$$

Scanned with OKEN Scanner

Rezultă din inegalitatea de mai sus că acest elipsoid cuprinde în întregime elipsoidul

$$y'By = k+2. \qquad \text{(XVII.6,5)}$$

In cazul particular cînd cei doi elipsoizi coincid, funcțiile de estimație corespunzătoare poartă numele de funcții de estimație eficiente.

Volumul închis de elipsoidul (XVII.6,5) este egal cu

$$V_B = \frac{\pi^{\frac{k}{2}}}{\Gamma\left(\frac{k}{2}+1\right)} \cdot \frac{(k+2)^{\frac{k}{2}}}{\sqrt{|B|}}.$$

La fel, volumul închis de elipsoidul (XVII.6,4) este

$$V_A = \frac{\pi^{\frac{k}{2}}}{\Gamma\left(\frac{k}{2}+1\right)} \cdot \frac{(k+2)^{\frac{k}{2}}}{\sqrt{|A|}}.$$

Avem

$$V_A \geqq V_B,$$

deci

$$\frac{V_B}{V_A} = \frac{\sqrt{|A|}}{\sqrt{|B|}} \leqq 1.$$

Putem măsura eficiența funcțiilor de estimație prin raportul

$$e(\bar{\alpha}_1, \bar{\alpha}_2, \ldots, \bar{\alpha}_k) = \frac{|A|}{|B|} \leqq 1.$$

Cînd acest raport este egal cu unu, vom fi în cazul funcțiilor de estimație eficientă.

Dacă limita lui, pentru $n \to \infty$, este egală cu unu, avem o eficiență asimptotică.

In sfîrșit, dacă ținem seama de proprietățile arătate în paragraful precedent, se vede că dacă ecuația verosimilității maxime admite o soluție, în condițiile arătate în acel paragraf, acea soluție este eficient asimptotică.

Capitolul XVIII

Teoria estimației pentru familiile de distribuții frecvent întîlnite în aplicații

§ 1. Schema lui Bernoulli

Să considerăm schema urnei lui Bernoulli cu două valori 1 și 0. Fie p probabilitatea ca să avem $X_n=1$ și q probabilitatea egalității $X_n=0$, probele $X_n(n=1, 2, ..)$ fiind presupuse independente. După n probe am obținut frecvența

$$f_n=\frac{\nu}{n}.$$

Ne punem problema estimării probabilității p.
Să considerăm funcția de estimație

$$\bar{\alpha}(\nu)=\frac{\nu}{n}.$$

Atunci

$$M(\bar{\alpha})=p,$$

$$D^2(\bar{\alpha})=\frac{pq}{n};$$

valoarea $\frac{\nu}{n}$ este deci o funcție de estimație absolut corectă.
Vom dovedi că această funcție de estimație este și eficientă.
Intr-adevăr, din teorema minimului dispersiei știm că, oricare ar fi funcția de estimație corectă $\bar{\beta}(\nu)$, există inegalitatea

$$D^2(\bar{\beta})\geqq\frac{1}{n\sum_{i=1}^{n}\left(\frac{\partial \ln p_i}{\partial p}\right)^2 p_i(\alpha)}.$$

In cazul nostru, parametrul α fiind egal cu p,

$$p_1(p)=p,$$

$$p_2(p)=1-p.$$

$$\sum_{i=1}^{n}\left(\frac{\partial \ln p_i}{\partial p_i}\right)^2 p_i=p\frac{1}{p^2}+(1-p)\frac{1}{(1-p)^2}=\frac{1}{pq}.$$

Prin urmare

$$D^2(\bar{\beta}) \geqq \frac{pq}{n}.$$

Se vede că minimul este atins pentru $\frac{\nu}{n}$, care este o funcţie de estimaţie eficientă.

La acelaşi rezultat ajungem aplicînd metoda verosimilităţii maxime

$$\frac{\partial \ln (p^{\alpha} q^{n-\alpha})}{\partial p} = \frac{\alpha}{p} - \frac{n-\alpha}{q}.$$

Egalînd această valoare cu zero, obţinem $p = \frac{\alpha}{n}$.

La acelaşi rezultat ajungem dacă, în loc de o singură serie de observaţii, considerăm un număr k de serii de observaţii independente, care ne-au dat respectiv frecvenţele

$$\frac{\nu_1}{n}, \quad \frac{\nu_2}{n}, \ldots, \frac{\nu_k}{n}.$$

Probabilitatea ca să obţinem frecvenţa $\frac{\nu_i}{n}$ este

$$C_n^{\nu_i} p^{\nu_i} q^{(n-\nu_i)} \qquad (i = 1, 2, \ldots, n).$$

Ecuaţia verosimilităţii maxime este

$$\frac{\partial \ln p^{\sum_1^k \nu_i} (1-p)^{kn - \sum_1^k \nu_i}}{\partial p} = 0.$$

Obţinem astfel funcţia de estimaţie

$$\bar{\alpha} = \frac{\sum_{i=1}^{k} \nu_i}{kn},$$

care este identică cu funcţia de estimaţie pe care am fi obţinut-o dacă cele k observaţii le-am fi contopit într-o singură serie de observaţii. Rezultă că nu se cîştigă în precizie dacă avem mai multe serii de observaţii.

Funcţia de estimaţie de mai sus este eficientă.

Scanned with OKEN Scanner

§ 2. Distribuția multinomială

Fie $p_i (i=1, 2, \ldots, k)$, probabilitățile respective de a obține într-o probă vectorii $\vec{e}_i (i=1, 2, \ldots, k)$

$$p_i \geqq 0, \quad \sum_{i=1}^{k} p_i = 1.$$

Într-o experiență de n probe independente, am obținut numerele

$$n_1, n_2, \ldots, n_{k-1}, n_k \qquad \left(\sum_{i=1}^{k} n_i = n\right)$$

reprezentînd de cîte ori au fost obținuți vectorii respectivi

$$e_1, e_2, \ldots, e_{k-1}, e_k.$$

Pe baza acestui rezultat, estimăm probabilitățile

$$p_1, p_2, \ldots, p_{k-1}.$$

Valoarea p_k rezultă ca o consecință din faptul că $p_1+p_2+\ldots+p_k=1$.
Funcția de verosimilitate este

$$P = \frac{n!}{n_1!\, n_2! \ldots n_k!} p_1^{n_1} p_2^{n_2} \ldots p_k^{n_k}.$$

Vom avea deci de rezolvat sistemul de ecuații

$$\frac{\partial \ln P}{\partial p_1} = 0, \quad \frac{\partial \ln P}{\partial p_2} = 0, \ldots, \frac{\partial \ln P}{\partial p_{k-1}} = 0,$$

care se poate scrie

$$\frac{n_1}{p_1} - \frac{n_k}{p_k} = 0, \quad \frac{n_2}{p_2} - \frac{n_k}{p_k} = 0, \ldots, \frac{n_{k-1}}{p_{k-1}} - \frac{n_k}{p_k} = 0.$$

Obținem astfel funcțiile de estimație

$$\bar{\alpha}_1 = \frac{n_1}{n}, \quad \bar{\alpha}_2 = \frac{n_2}{n}, \ldots, \bar{\alpha}_{k-1} = \frac{n_{k-1}}{n}.$$

Scanned with OKEN Scanner

Aceste funcţii de estimaţie sînt absolut corecte, deoarece avem (cap. XVI, § 3)

$$M(\bar{\alpha}_i)=p_i,$$

$$D^2(\bar{\alpha}_i)=\frac{p_i(1-p_i)}{n} \qquad (i=1,\ 2,\ \ldots,\ k-1).$$

După teoria generală, urmează că aceste estimaţii sînt asimptotic eficiente. Se poate însă arăta mai mult decît atît şi anume că sînt eficiente.

Elipsoidul de concentraţie minimă al dispersiei are drept coeficienţi ai necunoscutelor expresiile

$$n\sum_{h=1}^{k}\left(\frac{\partial \ln p_h}{\partial p_i}\right)^2 p_h=\frac{p_i+p_k}{p_ip_k}\,n \qquad (i=1,\ 2,\ \ldots,\ k-1),$$

$$n\sum\frac{\partial \ln p_h}{\partial p_i}\cdot\frac{\partial \ln p_h}{\partial p_j}\,p_h=\frac{1}{p_k}\,n \qquad (i,\ j=1,\ 2,\ \ldots,\ k-1,\ i\neq j),$$

Prin urmare, ecuaţia elipsoidului de concentraţie minimă al dispersiei este

$$\sum_{i=1}^{k-1}\frac{p_i+p_k}{p_i}(x_i-p_i)^2+2\sum_{\substack{i,j=1\\(i\neq j)}}^{k-1}(x_i-p_i)(x_i-p_j)=\frac{(k+1)\,p_k}{n}. \qquad \text{(XVIII.2,1)}$$

Pe de altă parte, din formulele

$$D^2(\bar{\alpha}_i)=\frac{p_i(1-p_i)}{n},\quad \rho(\bar{\alpha}_i,\ \bar{\alpha}_j)=-\frac{p_ip_j}{n},$$

rezultă că avem de considerat matricea

$$A=\left\|\begin{array}{cccc}
\frac{p_1(1-p_1)}{n} & \frac{-p_1p_2}{n} & \ldots & \frac{-p_1p_{k-1}}{n}\\
\frac{-p_2p_1}{n} & \frac{p_2(1-p_2)}{n} & \ldots & \frac{-p_2p_{k-1}}{n}\\
\ldots & \ldots & \ldots & \ldots\\
\frac{-p_{k-1}p_1}{n} & \frac{-p_{k-1}p_2}{n} & \ldots & \frac{p_{k-1}(1-p_{k-1})}{n}
\end{array}\right\|.$$

Un calcul elementar ne conduce la matricea inversă

$$A^{-1}=\left\|\begin{array}{cccc} \dfrac{n(p_1+p_k)}{p_1p_k} & \dfrac{n}{p_k} & \cdots & \dfrac{n}{p_k} \\ \dfrac{n}{p_k} & \dfrac{n(p_2+p_k)}{p_2p_k} & \cdots & \dfrac{n}{p_k} \\ \cdots & \cdots & \cdots & \cdots \\ \dfrac{n}{p_k} & \dfrac{n}{p_k} & \cdots & \dfrac{n(p_{k-1}+p_k)}{p_{k-1}p_k} \end{array}\right\|$$

şi deci la ecuaţia (XVIII.2,1).

Elipsoidul de concentraţie al funcţiilor de estimaţie $\overline{\alpha}_1, \overline{\alpha}_2, \ldots, \overline{\alpha}_{k-1}$ coincide cu elipsoidul de concentraţie minimă al dispersiei.

Prin urmare, funcţiile de estimaţie găsite sînt eficiente.

§ 3. Distribuţia Poisson

Funcţia de distribuţie este

$$f(x, \lambda)=\frac{\lambda^x e^{-\lambda}}{x!},$$

unde λ este parametrul pe care ne propunem să-l estimăm în baza faptului că în n experienţe independente am obţinut succesiv valorile

$$x_1, x_2, \ldots, x_n.$$

Funcţia de verosimilitate maximă este

$$P=e^{-n\lambda}\frac{\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!}.$$

Ecuaţia de verosimilitate maximă

$$\frac{\partial \ln P}{\partial \lambda}=0$$

ne dă funcţia de estimaţie

$$\overline{\lambda}=\frac{\sum_{i=1}^{n} x_i}{n}.$$

Pe de altă parte, minimul dispersiei pentru funcțiile de estimație ne este dat de expresia

$$\frac{1}{n\sum_{\lambda=0}^{\infty}\left[\frac{\partial \ln f(x;\lambda)}{\partial\lambda}\right]^2 f(x;\lambda)}.$$

Avem însă

$$n\sum_{\lambda=0}^{\infty}\left[\frac{\partial \ln f(x,\lambda)}{\partial\lambda}\right]^2 f(x,\lambda)=n\sum_{x=0}^{\infty}\left(\frac{x}{\lambda}-1\right)^2\frac{\lambda^x e^{-x}}{x!}=\frac{n}{\lambda}.$$

Minimul dispersiei este prin urmare $\frac{\lambda}{n}$.

Funcția de estimație atinge acest minim, deoarece avem

$$D^2(\lambda)=\frac{\lambda}{n}$$

și $\overline{\lambda}$ este prin urmare o funcție de estimație eficientă.

§ 4. Legea normală. Estimația mediei

Să considerăm familia de distribuții normale cu funcția de frecvență

$$f(x,\ m)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-m)}{2\sigma^2}},$$

unde σ este cunoscut, iar m un parametru variabil, care urmează să fie estimat pe baza a n observații independente.

Să vedem dacă există o statistică exhaustivă. Pentru aceasta, vom aplica mai multe metode.

1. *Metoda bazată pe forma particulară a funcției de frecvență.* Am văzut că pentru existența unei statistici exhaustive este necesar și suficient ca $f(x,\ m)$ să aibă forma particulară

$$\ln f(x,\ m)=A_1(m)\,\psi_1(x)+A_0(m)+\psi_0(x)$$

și că, în acest caz, statistica exhaustivă este dată de funcția

$$\overline{\alpha}(x)=\frac{\sum_{i=1}^{n}\psi_i(x_i)}{n}.$$

In cazul de față

$$\ln f(x,\ m) = \frac{m}{\sigma^2} x + \frac{m^2}{2\sigma^2} - \frac{x^2}{2\sigma^2},$$

deci

$$A_1(m) = \frac{m}{\sigma^2}, \quad A_0(m) = \frac{m}{2\sigma^2},$$

$$\psi_1(x) = x, \quad \psi_0(x) = -\frac{x^2}{2\sigma^2}.$$

Statistica exhaustivă este

$$\overline{\alpha}(\vec{x}) = \frac{\sum_{i=1}^{n} x_i}{n}.$$

2. *Metoda bazată pe determinarea unei statistici necesare și suficiente.* Am arătat că obținem o statistică necesară și suficientă cu ajutorul formulei

$$g(x,\ m) = \ln f(x,\ m) - \ln f(x,\ m_0),$$

unde m_0 este o valoare particulară dată parametrului m. In cazul nostru, obținem

$$g(x,\ m) = -\frac{(x-m)^2}{2\sigma^2} + \frac{(x-m_0)^2}{2\sigma^2} = \frac{m-m_0}{\sigma^2} x - \frac{m^2-m_0^2}{2\sigma^2}.$$

Statistica necesară și suficientă este

$$\sum_{i=1}^{n} g(x_i,\ m) = \frac{m-m_0}{\sigma^2} \sum_{i=1}^{n} x_i - n\,\frac{m^2-m_0^2}{2\sigma^2},$$

care este echivalentă cu $\sum_{i=1}^{n} x_i$. Putem deci lua ca statistică exhaustivă aceeași funcție

$$\alpha(\vec{x}) = \frac{\sum_{i=1}^{n} x_i}{n}$$

ca mai sus. Intrucît

$$M\left(\frac{\sum_{i=1}^{n} x_i}{n}\right) = m,$$

statistica exhaustivă $\overline{\alpha}(x)$ ne poate furniza o funcţie de estimaţie eficientă sau numai o funcţie de estimaţie suficientă.

Insă

$$D^2[\overline{\alpha}(x)] = D^2\left(\frac{\sum_{i=1}^{n} x_i}{n}\right) = \frac{1}{n^2}\sum_{i=1}^{n} D^2(x_i) = \frac{\sigma^2}{n},$$

$$\overline{\alpha}(\vec{x}) = \frac{\sum_{i=1}^{n} x_i}{n}$$

este, prin urmare, o estimaţie absolut corectă.

Aplicînd atunci teorema minimului dispersiei, rezultă că pentru orice funcţie de estimaţie absolut corectă $\overline{\beta}(x)$, avem

$$D^2(\overline{\beta}(x)) \geqq \frac{1}{\int_{-\infty}^{+\infty}\left(\frac{\partial \ln f(x, m)}{\partial m}\right)^2 f(x, m)\,dx},$$

sau, în cazul nostru, efectuînd calculele,

$$D^2(\overline{\beta}(x)) \geqq \frac{\sigma^2}{n}.$$

Se vede că funcţia de estimaţie absolut corectă $\overline{\alpha}(x)$ atinge minimul $\frac{\sum_{i=1}^{n} x_i}{n}$ şi este prin urmare o funcţie de estimaţie eficientă.

3. *Metoda bazată pe principiul verosimilităţii maxime.* Probabilitatea obţinerii valorilor $x_1, x_2, \ldots, x_n$ este

$$P = \frac{1}{\sigma^n (2\pi)^{\frac{n}{2}}} e^{-\frac{\sum_{i=1}^{n}(x_i - m)^2}{2\sigma^2}}.$$

Din ecuația verosimilității maxime $\ln P=0$, rezultă

$$\sum_{i=1}^{n}(x_i-m)=0$$

sau

$$\frac{\sum_{i=1}^{n}x_i}{n}=m.$$

Ajungem la aceeași funcție de estimație care, conform teoriei generale, trebuie să fie eficientă.

4. *O metodă geometrică.* Vom interpreta geometric rezultatele obținute, considerînd pentru simplificare $n=3$. Punctul $P(x_1, x_2, \ldots, x_n)$ din spațiul V^3 are densitatea de distribuție

$$\frac{1}{(2\pi)^{\frac{3}{2}}\sigma^3}e^{-\frac{\sum_{i=1}^{n}(x_i-m)^2}{2\sigma^2}}.$$

Prin urmare $P(x_1, x_2, x_3)$ urmează o lege normală cu trei dimensiuni. Fie $Ox_1\ x_2\ x_3$ un triedru de referință ortogonal. Proiecția punctului $P(x_1, x_2, x_3)$ pe dreapta D, avînd ecuația $x_1=x_2=x_3$, este punctul $L\left(\frac{1}{3}\sum_{i=1}^{3}x_i,\ \frac{1}{3}\sum_{i=1}^{3}x_i,\ \frac{1}{3}\sum_{i=1}^{3}x_i\right)$.

Să considerăm acum pe dreapta D punctul $N(m, m, m)$.

In triunghiul dreptunghic PNL avem

$$\overline{PN}^2=\overline{NL}^2+\overline{LP}^2,$$

sau

$$\sum_{i=1}^{n}(x_i-m)^2=3\left(x-\frac{1}{3}\sum_{i=1}^{3}x_i\right)^2+\sum_{i=1}^{3}\left(x_i-\frac{1}{3}\sum_{i=1}^{3}x_i\right)^2.$$

Reperarea punctului $P(x_1, x_2, x_3)$ poate fi astfel făcută cu ajutorul segmentului $\overline{NP}=\sqrt{3}\left(\frac{1}{3}\sum_{i=1}^{n}x_i-m\right)$, al segmentului $\overline{LP}$, pe care-l notăm cu $\overline{\sigma}$ și al unui coeficient $\overline{\alpha}$ necesar pentru a determina poziția punctului P, în planul π, perpendicular pe D, dus prin P. Prin urmare, $\overline{\sigma}$ și $\overline{\alpha}$ constituie coordonatele polare ale punctului P în planul π.

Scanned with OKEN Scanner

Cu aceste elemente, probabilitatea punctului P se poate scrie

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{\overline{NL}^2}{2\sigma^2}} \frac{d(\overline{LN})}{\sigma} e^{-\frac{\bar{\sigma}^2}{2\sigma^2}} \frac{\bar{\sigma}\, d\bar{\sigma}}{\sigma^2} \frac{d\alpha}{2\pi},$$

care pune în evidență următoarele fapte:

1) punctul P urmează pe D o lege normală (m, σ);

2) $\bar{\sigma}$ urmează o lege χ cu două grade de libertate;

3) α are o repartiție uniformă.

Legea punctului $P(x_1, x_2, \ldots, x_n)$ poate fi scrisă

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{\overline{NL}^2}{2\sigma^2}} \frac{d(\overline{NL})}{\sigma} \cdot \frac{2^{\frac{n-3}{2}}}{\Gamma\left(\frac{n-1}{2}\right)} e^{-\frac{\bar{\sigma}^2}{2\sigma^2}} \cdot \frac{\bar{\sigma}^{n-2} d\bar{\sigma}}{\sigma^{n-1}} \frac{ds}{s}.$$

In acest caz, punctul $L\left(\frac{1}{n}\sum_{i=1}^{n} x_i, \frac{1}{n}\sum_{i=1}^{n} x_i, \ldots, \frac{1}{n}\sum_{i=1}^{n} x_i\right)$ reprezintă proiecția lui P pe dreapta D, având ecuațiile

$$x_1 = x_2 = \ldots = x_n.$$

El se supune pe dreapta D unei legi normale (m, σ).

Pe sfera $(n-1)$-dimensională, cu centrul în L și raza LP, poziția punctului P este fixată de raza $LP = \bar{\sigma}$ care urmează o lege χ cu $n-1$ grade de libertate și de $n-2$ coeficienți unghiulari, care dau o lege de probabilitate uniformă pentru poziția punctului P pe sferă.

Forma particulară a acestei distribuții pune în evidență statistica exhaustivă existentă.

§ 5. Legea normală. Estimația valorii medii pătratice

Să considerăm familia de distribuții normale cu funcția de frecvență

$$f(x, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}},$$

unde media m este cunoscută, iar σ este un parametru necunoscut, care urmează să fie estimat în baza observațiilor independente $x_1, x_2, \ldots, x_n$.

Să cercetăm existența unei statistici exhaustive.

1. *Metoda bazată pe forma particulară a funcției de frecvență.* Existența unei statistici exhaustive implică pentru $f(x, \sigma)$ forma particulară

$$\ln f(x, \sigma) = A_1(\sigma)\,\psi_1(x) + A_0(\sigma) + \psi_0(x),$$

statistica exahaustivă fiind dată de funcția

$$\alpha(\vec{x}) = \frac{\sum_{i=1}^{n} \psi_1(x_i)}{n}.$$

In cazul de față

$$\ln f(x, \sigma) = -\ln \sigma - \ln \sqrt{2\pi} - \frac{(x-m)^2}{2\sigma^2}.$$

Prin urmare

$$A_1(\sigma) = -\frac{1}{2\sigma^2}, \quad A_0(\sigma) = \ln \sigma - \ln \sqrt{2\pi},$$

$$\psi_1(x) = (x-m)^2, \quad \psi_0(x) = 0.$$

Avem deci statistica exhaustivă

$$\alpha(\vec{x}) = \frac{\sum_{i=1}^{n} (x_i - m)^2}{n}.$$

2. *Metoda bazată pe determinarea unei statistici necesare și suficiente.* O statistică necesară si suficientă se poate obține aplicînd relația

$$g(x, \sigma) = \ln f(x, \sigma) - \ln f(x, \sigma_0),$$

unde σ_0 este o valoare particulară dată parametrului σ. Avem

$$g(x, \sigma) = \frac{(x-m)^2}{2\sigma^2} + \frac{(x-m)^2}{2\sigma_0^2} = \frac{1}{2}\left[\frac{1}{\sigma_0^2} - \frac{1}{\sigma^2}\right](x-m)^2,$$

$$\sum_{i=1}^{n} g(x_i, \sigma) = \frac{1}{2}\left(\frac{1}{\sigma_0^2} - \frac{1}{\sigma}\right)\sum_{i=1}^{n} (x_i - m)^2.$$

O statistică necesară și suficientă echivalentă este

$$\alpha(\vec{x}) = \frac{\sum_{i=1}^{n} (x_i - m)^2}{n}.$$

Pentru obținerea unei funcții de estimație absolut corecte trebuie deci să considerăm o funcție de $\bar{\alpha}$, $\bar{\beta}=\varphi(\bar{\alpha})$ astfel încît

$$M(\bar{\beta})=\sigma, \quad D^2(\bar{\beta})\underset{n\to\infty}{\to}0.$$

a) Să considerăm pentru $\bar{\beta}$ funcția

$$\bar{\beta}=\sqrt{\frac{\sum_{i=1}^{n}(x_i-m)^2}{n}}.$$

Variabila (x_i-m) urmează o lege de distribuție normală $(0, \sigma)$. Pe de altă parte variabilele (x_i-m) $(i=1, 2,..., n)$ sînt independente. Legea de probabilitate a lui $\bar{\beta}$ este

$$\frac{2\left(\frac{n}{2}\right)^{\frac{n}{2}}}{\sigma^n\,\Gamma\left(\frac{n}{2}\right)}x^{n-1}\cdot e^{-\frac{n}{2\sigma^2}x^2}.$$

Efectuînd calculele, rezultă

$$M(\bar{\beta})=\sqrt{\frac{n}{2}}\,\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n+1}{2}\right)}\sigma.$$

Considerînd atunci expresia

$$\sigma(\vec{x})=\sqrt{\frac{n}{2}}\,\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n+1}{2}\right)}\sqrt{\frac{\sum_{i=1}^{n}(x_i-m)^2}{n}}$$

sau

$$\bar{\sigma}(\vec{x})=\frac{1}{\sqrt{2}}\,\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n+1}{2}\right)}\sqrt{\sum_{i=1}^{n}(x_i-m)^2},$$

obținem

$$M(\bar{\sigma})=\sigma,$$

$$D^2(\bar{\sigma})=\sigma^2\left[\frac{n}{2}\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma^2\left(\frac{n+1}{2}\right)}-1\right].$$

Asimptotic avem

$$\Gamma(x+1)=x^x e^{-x}\sqrt{2\pi x},$$

de unde rezultă

$$D^2(\bar{\sigma})=\frac{\sigma^2}{2n}+O\left(\frac{1}{n^2}\right),$$

$$D^2(\bar{\sigma})\to 0.$$

$$n\to\infty$$

Prin urmare $\bar{\sigma}(\vec{x})$ este o funcție de estimație absolut corectă a lui σ.

Se vede ușor că $\bar{\sigma}(\vec{x})$ se poate scrie și sub forma

$$\bar{\sigma}(\vec{x})=\frac{2\cdot4\cdot6\ldots(n-2)}{1\cdot3\cdot5\ldots(n-1)}\sqrt{\frac{2}{\pi}\sum_{i=1}^{n}(x_i-m)^2}\qquad\text{pentru } n \text{ par},$$

$$\bar{\sigma}(\vec{x})=\frac{1\cdot3\cdot5\ldots(n-2)}{2\cdot4\cdot6\ldots(n-1)}\sqrt{\frac{\pi}{2}\sum_{i=1}^{n}(x_i-m)^2}\qquad\text{pentru } n \text{ impar}.$$

Teorema minimului dispersiei ne arată că

$$D^2(\bar{\sigma})\geqq\frac{1}{nM\left(\frac{\partial \ln f}{\partial\sigma}\right)^2}.$$

Însă

$$M\left(\frac{\partial \ln f}{\partial\sigma}\right)^2=\int_{-\infty}^{+\infty}\left[\frac{(x-m)^2}{\sigma^3}-\frac{1}{\sigma}\right]\cdot f(x,\ \sigma)\,d\sigma=\frac{2}{\sigma^2}.$$

Prin urmare

$$D^2(\sigma)\geqq\frac{\sigma^2}{2n}.$$

Calculînd valoarea lui $D^2(\bar{\sigma})$, am văzut că diferă de $\frac{\sigma^2}{2n}$. Prin urmare, $\bar{\sigma}$ nu este o funcție de estimație eficientă a lui σ, ci numai o funcție de estimație suficientă, în virtutea existenței unei statistici exhaustive.

Eficiența $e_n(\bar{\sigma})$ este dată de formula

$$e_n(\bar{\sigma}) = \frac{\frac{1}{2n}}{\frac{n}{2}\frac{\Gamma^2\left(\frac{n}{2}\right)}{\Gamma^2\left(\frac{n+1}{2}\right)} - 1}.$$

Aşadar, obţinem

$$\lim_{n\to\infty} e_n(\bar{\sigma}) = 1;$$

prin urmare $\bar{\sigma}$ este o funcție de estimație eficient asimptotică.

b) Știm că există o infinitate de funcții de estimație absolut corecte ale lui σ. Să considerăm o altă funcție de estimație absolut corectă, pornind de la dispersia empirică

$$\bar{\sigma}^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})}{n}.$$

Am arătat că $\bar{\sigma}$ are densitatea de distribuție

$$f(x) = \frac{2n^{\frac{n-1}{2}}}{2^{\frac{n-1}{2}}\sigma^{n-1}\Gamma\left(\frac{n-1}{2}\right)}\sigma^{n-2}e^{-\frac{nx^2}{2\sigma^2}}.$$

Efectuînd calculele rezultă că

$$M(\bar{\sigma}) = \sqrt{\frac{n}{2}}\frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}\sigma.$$

Rezultă că, dacă vom considera funcția

$$\bar{\delta}(\vec{x}) = \sqrt{\frac{n}{2}} \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}},$$

atunci

$$M(\bar{\sigma}) = \sigma.$$

Efectuînd calculele, obținem

$$D^2(\bar{\delta}) = \left[\frac{n-1}{2} \frac{\Gamma^2\left(\frac{n-1}{2}\right)}{\Gamma^2\left(\frac{n}{2}\right)} - 1 \right] \sigma^2,$$

$$D^2(\bar{\delta}) = \frac{\sigma^2}{2n} + O\left(\frac{1}{n^2}\right).$$

Prin urmare și $\bar{\delta}(\vec{x})$ este o funcție de estimație absolut corectă. Eficiența $e_n(\bar{\delta})$ este dată de relația

$$e_n(\delta) = \frac{\frac{1}{2n}}{\frac{n-1}{2} \frac{\Gamma^2\left(\frac{n-1}{2}\right)}{\Gamma^2\left(\frac{n}{2}\right)} - 1}.$$

Cînd $n \to \infty$, $e_n(\bar{\delta}) \to 1$. Prin urmare $\bar{\delta}$ este de asemenea o funcție de estimație eficient asimptotică. Valoarea $\bar{\delta}$ încetează însă de a mai fi o funcție de estimație suficientă, deoarece nu este exprimată numai în funcție de statistica exhaustivă $\bar{\alpha}(\vec{x})$. $\bar{\delta}(x)$, după cum vedem, face abstracție de faptul că ne este cunoscută adevărata valoare a lui m, pe cînd $\bar{\sigma}(\vec{x})$ cuprinde pe m. De aceea $\bar{\sigma}(\vec{x})$ constituie o funcție de estimație mai precisă decît $\bar{\delta}(\vec{x})$.

Scanned with OKEN Scanner

Intr-adevăr, notînd respectiv cu $e_n(\bar{\sigma})$ şi $e_n(\bar{\delta})$ eficienţa lui $\bar{\sigma}$ şi $\bar{\delta}$ pentru grupul informativ de valori $x_1, x_2, \ldots, x_n$, avem

$$e_2(\bar{\sigma}) = \frac{\pi}{4(4-\pi)} = 0{,}915\,1, \quad e_2(\bar{\delta}) = \frac{1}{2(\pi-2)} = 0{,}438\,0,$$

$$e_3(\bar{\sigma}) = \frac{4}{3(3\pi-8)} = 0{,}935\,8, \quad e_3(\bar{\delta}) = \frac{\pi}{6(4-\pi)} = 0{,}610\,0.$$

3. *Metoda bazată pe principiul verosimilităţii maxime.* Probabilitatea de a obţine valorile $x_1, x_2, \ldots, x_n$ este

$$P = \frac{1}{\sigma^n (2\pi)^{\frac{n}{2}}} e^{-\frac{\sum_{i=1}^{n} (x_i - m)^2}{2\sigma^2}};$$

$$\ln P = -n \ln \sigma - \frac{n}{2} \ln 2\pi - \frac{\sum_{i=1}^{n} (x_i - m)^2}{2\sigma^2}.$$

Ecuaţia verosimilităţii maxime

$$\frac{\partial \ln P}{\partial \sigma} = 0$$

ne conduce la soluţia

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - m)^2}{n}}.$$

Conform celor stabilite, dacă există o funcţie de estimaţie suficientă, ea este funcţie de valoarea mai sus obţinută, ceea ce am constatat direct cu funcţia de estimaţie $\bar{\sigma}(\vec{x})$.

§ 6. Legea normală. Estimaţia dispersiei

Să plecăm tot de la familia de distribuţii normale

$$f(x, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}},$$

unde media m este cunoscută şi σ este un parametru necunoscut.

Am procedat mai sus la estimația lui σ. Ne propunem acum să găsim o funcție de estimație pentru σ^2. Am văzut că

$$\bar{\sigma}(\vec{x}) = \frac{1}{\sqrt{2}} \cdot \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n+1}{2}\right)} \sqrt{\sum_{i=1}^{n} (x_i - m)^2}$$

este o funcție de estimație suficientă. Conform teoriei generale, rezultă că putem determina o funcție $\beta(\sigma)$ astfel încît pentru β să găsim o funcție de estimație suficientă. Așadar, dacă

$$\ln f(x, \sigma) = A_1(\sigma)\,\Psi_1(x) + A_0(\sigma) + \Psi_0(x),$$

adică avem îndeplinită condiția necesară și suficientă pentru existența unei funcții de estimație suficiente, am văzut că substituția

$$\theta = -\frac{A_0'(\sigma)}{A_1'(\sigma)}$$

ne dă posibilitatea obținerii unei funcții de estimație eficiente pentru θ.

In cazul nostru

$$\ln f(x, \sigma) = -\frac{(x-m)^2}{2\sigma^2} - \ln\sigma - \ln\sqrt{2\pi}.$$

Prin urmare

$$A_0(\sigma) = -\ln\sigma - \ln\sqrt{2\pi},$$

$$A_1(\sigma) = -\frac{1}{2\sigma^2},$$

$$\theta = \sigma^2.$$

Pentru σ^2 putem deci găsi o funcție de estimație eficientă. Metoda verosimilității maxime ne dă în acest caz estimația căutată.

Am arătat mai sus că

$$\ln P = -n\ln\sigma - \frac{n}{2}\ln\sqrt{2\pi} - \frac{\sum_{i=1}^{n}(x_i - m)^2}{2\sigma^2}.$$

Făcînd substituția $\theta = \sigma^2$ rezultă

$$\ln P = -\frac{n}{2}\ln\theta - \frac{n}{2}\ln\sqrt{2\pi} - \frac{\sum_{i=1}^{n}(x_i - m)^2}{2\theta}.$$

Scanned with OKEN Scanner

Din ecuația verosimilității maxime obținem atunci

$$\theta = \frac{\sum_{i=1}^{n} (x_i - m)^2}{n},$$

care constituie o funcție de estimație eficientă pentru σ^2.

Putem verifica cu ajutorul teoremei minimului dispersiei că într-adevăr expresia de mai sus constituie o funcție de estimație eficientă. Atunci

$$\left(\frac{\partial \ln f}{\partial \sigma}\right)^2 = \int_{-\infty}^{+\infty} \left[\frac{(x-m)^2}{2\sigma^4} - \frac{1}{2\sigma^4}\right]^2 f dx = \frac{1}{2\sigma^4}.$$

Deci dispersia oricărei estimații absolut corectè a lui σ^2 este mai mare sau cel mult egală cu $\frac{2\sigma^4}{n}$.

Insă

$$D^2\left(\frac{\sum_{i=1}^{n}(x_i - m)^2}{n}\right) = \frac{2\sigma^4}{n}.$$

Minimul este atins, verificînd astfel că $\frac{\sum_{i=1}^{n} (x_i - m)^2}{n}$ este o funcție de estimație eficientă.

Am văzut că

$$\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

este o funcție de estimație absolut corectă a lui σ^2. Intr-adevăr,

$$M\left(\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}\right) = \sigma^2,$$

$$D^2\left(\frac{\sum_{i=1}^{n} (x_i - \bar{x})}{n-1}\right) = \frac{2\sigma^2}{n-1}.$$

Eficiența acestei funcții de estimație este

$$e_n\left(\frac{\sum_{i=1}^{n}(x_i-x)^2}{n-1}\right)=\frac{n-1}{n}.$$

Cînd $n\to\infty$, eficiența e_n tinde către unu. Deci $\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$ reprezintă o funcție de estimație asimptotic eficientă.

§ 7. Legea normală. Mediana ca funcție de estimație

Pentru distribuția normală

$$f(x,\ m,\ \sigma)=\frac{1}{\sigma\sqrt{2\pi}}\,e^{-\frac{(x-m)^2}{2\sigma^2}}$$

mediana este egală cu m.
Este natural să ne propunem estimarea lui m cu ajutorul medianei de selecție din n probe independente, care au dat rezultatul

$$x_1,\quad x_2,\ldots,x_n.$$

Fie ξ_n mediana acestor valori.
La limită ξ_n urmează o distribuție normală, cu următoarele valori tipice

$$M(\xi_n)=m,$$

$$D^2(\xi_n)=\sigma\sqrt{\frac{\pi}{2n}}.$$

Aceste formule se pot ușor demonstra.
Să considerăm o selecție de mărime $2n+1$

$$x_1,\quad x_2,\ldots,x_{2n+1}.$$

In acest caz mediana selecției este x_n.

Să presupunem că funcția de repartiție a distribuției teoretice este $F(x)$ și că există o densitate de probabilitate $F'(x)=f(x)$. Mediana fiind notată prin m, avem

$$F(m)=\frac{1}{2}.$$

Probabilitatea ca să existe relația $x<x_n<x+dx$ este

$$\varphi(x)\,dx=\frac{(2n+1)!}{(n!)\sqrt{n}}[F(x)]^n[1-F(x)]^n f(x)\,dx.$$

Să considerăm variabila $y=\sqrt{n}\,(x-m)$.

Din formula precedentă rezultă că funcția de frecvență corespunzătoare variabilei y este

$$\frac{(2n+1)!}{(n!)^2\sqrt{n}}\left[F\left(m+\frac{y}{\sqrt{n}}\right)\right]^n\left[1-F\left(m+\frac{y}{\sqrt{n}}\right)\right]^n f\left(m+\frac{y}{n}\right).$$

Ținînd seamă de formula lui Stirling și de dezvoltarea

$$F\left(m+\frac{y}{\sqrt{n}}\right)=\frac{1}{2}+\frac{y}{\sqrt{n}}f(m)+\frac{y^2}{2n}f'(m)+O(n^{-\frac{3}{2}}),$$

probabilitatea de mai sus devine asimptotic

$$\frac{2f(m)}{\sqrt{\pi}}e^{-4y^2f(m)}.$$

Rezultă deci că într-o selecție de mărime $(2n+1)$ mediana selecției este distribuită normal cu media în mediana distribuției teoretice și cu dispersia $\sigma^2=\frac{1}{8f^2(m)}$.

In cazul în care $f(x)$ reprezintă funcția de frecvență a distribuției normale și ne referim la o selecție de mărime n, ajungem la formulele indicate.

Funcția de estimație ξ_n, dată de mediana experimentală, este deci absolut corectă. Ea însă nu este eficientă. Eficiența ei este egală cu

$$e(\xi_n)=\frac{2}{\pi}=0{,}637\ldots$$

Prin urmare o sută de observații ale medianei ne dau pentru estimarea lui m aceeași precizie ca șaizeci și patru observații ale mediei aritmetice.

§ 8. Legea normală. Estimația valorii medii și a valorii medii pătratice

Cazul pe care-l considerăm se referă la estimația ambilor parametri m și σ care intervin în familia de distribuții normale

$$f(x,\ m,\ \sigma)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-m)^2}{2\sigma^2}}.$$

Să examinăm existența statisticilor exhaustive.

1. *Metoda bazată pe forma particulară a funcției de frecvență.* Am demonstrat că, pentru ca să avem o statistică exhaustivă de ordin minim, este necesar și suficient ca

$$\ln f(x,\ m,\ \sigma)=\sum_{k=1}^{m}A_k(m,\ \sigma)\,\Psi_k(x)+A_0(m,\ \sigma)+\Psi_0(x).$$

În cazul nostru,

$$\ln f(x,\ m,\ \sigma)=-\frac{x}{2\sigma^2}+\frac{mx}{\sigma^2}-\frac{m^2}{2\sigma^2}-\ln\sigma-\ln\sqrt{2\pi},$$

$$A_1(m,\ \sigma)=-\frac{1}{2\sigma^2},\quad \Psi_1(x)=x^2,$$

$$A_2(m,\ \sigma)=\frac{m}{\sigma^2},\quad \Psi_2(x)=x,$$

$$A_0(m,\ \sigma)=-\frac{m^2}{2\sigma^2}-\ln\sigma-\ln\sqrt{2\pi},\quad \Psi_0(x)=0.$$

Obținem așadar o statistică exhaustivă de ordin minim doi, dată de funcțiile

$$\sum_{i=1}^{n}x_i,\quad \sum_{i=1}^{n}x_i^2.$$

2. *Metoda bazată pe determinarea unei statistici necesare și suficiente.* Obținem o statistică necesară și suficientă cu ajutorul formulei

$$g(x,\ m,\ \sigma)=\ln f(x,\ m,\ \sigma)-\ln f(x,\ m_0,\ \sigma_0),$$

unde m_0 și σ_0 reprezintă valori particulare date parametrilor m și σ.

Deci

$$g(x,\ m,\ \sigma)=\frac{1}{2}\left(\frac{1}{\sigma_0^2}-\frac{1}{\sigma^2}\right)x^2+\left(\frac{m}{\sigma^2}-\frac{m_0}{\sigma_0^2}\right)x,$$

$$\sum_{i=1}^{n}g(x_i,\ m,\ \sigma)=\frac{1}{2}\left(\frac{1}{\sigma_0^2}-\frac{1}{\sigma^2}\right)\sum_{i=1}^{n}x_i^2+\left(\frac{m}{\sigma^2}-\frac{m_0}{\sigma_0^2}\right)\sum_{i=1}^{n}x_i.$$

Această statistică necesară și suficientă este echivalentă cu statistica exhaustivă

$$\sum_{i=1}^{n}x_i,\quad \sum_{i=1}^{n}x_i^2,$$

găsită prin metoda 1.

Pentru rezolvarea problemei, va trebui deci să alegem două funcții de estimație $\bar{\alpha}(\vec{x})$ și $\bar{\beta}(\vec{x})$ de forma

$$\bar{\alpha}(\vec{x})=F\left(\sum_{i=1}^{n}x_i,\quad \sum_{i=1}^{n}x_i^2\right),$$

$$\bar{\beta}(\vec{x})=G\left(\sum_{i=1}^{n}x_i,\quad \sum_{i=1}^{n}x_i^2\right),$$

astfel încît $\bar{\alpha}(\vec{x})$ și $\bar{\beta}(\vec{x})$ să reprezinte respectiv funcții de estimație absolut corecte pentru m și σ.

Estimația lui m. Fie

$$\bar{\alpha}(\vec{x})=\frac{\sum_{i=1}^{n}x_i}{n}.$$

Atunci

$$M(\bar{\alpha})=m,\quad D^2(\bar{\alpha})=\frac{\sigma^2}{n}$$

pentru orice m și σ. Prin urmare $\bar{\alpha}(\vec{x})$ reprezintă o funcție de estimație absolut corectă a lui m.

Estimația lui σ. Am văzut că

$$\bar{\sigma}^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n} = \frac{\sum_{i=1}^{n}x_i^2}{n} - \left(\frac{\sum_{i=1}^{n}x_i}{n}\right)^2$$

se bucură de proprietatea că

$$M(\bar{\sigma}) = \frac{\sqrt{2}}{\sqrt{n}} \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \sigma. \qquad \text{(XVIII.8,1)}$$

Putem deci alege pentru estimația lui σ o funcție de $\sum_{i=1}^{n}x_i$ și $\sum_{i=1}^{n}x_i^2$

$$\bar{\beta}(x) = \sqrt{\frac{1}{2}} \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \sqrt{\sum_{k=1}^{n}(x_k-\bar{x})^2}$$

sau

$$\bar{\beta}(\vec{x}) = \frac{1\cdot3\cdot5\ldots(n-3)}{2\cdot4\cdot6\ldots(n-2)} \sqrt{\frac{\pi}{2}\sum_{i=1}^{n}(x_i-\bar{x})^2} \qquad \text{pentru } n \text{ par}$$

$$\bar{\beta}(\vec{x}) = \frac{2\cdot4\cdot6\ldots(n-2)}{1\cdot3\cdot5\ldots(n-3)} \sqrt{\frac{2}{\pi}\sum_{i=1}^{n}(x_i-\bar{x})^2} \qquad \text{pentru } n \text{ impar.}$$

Atunci

$$D^2(\bar{\beta}) = \left[\frac{n-1}{2} \cdot \frac{\Gamma^2\left(\frac{n-1}{2}\right)}{\Gamma^2\left(\frac{n}{2}\right)} - 1\right]\sigma^2,$$

$$D^2(\bar{\beta}) = \frac{\sigma^2}{2n} + O\left(\frac{1}{n^2}\right);$$

$\bar{\beta}(\vec{x})$ este deci o estimație absolut corectă a lui σ.
Am văzut pe de altă parte la teoria selecției că cele două variabile $\bar{\alpha}(\vec{x})$ și $\bar{\beta}(\vec{x})$ sînt independente între ele.

Estimaţia lui σ^2. Putem alege pentru estimaţia lui σ^2 funcţia de $\sum_{i=1}^{n} x_i$ şi $\sum_{i=1}^{n} x_i^2$

$$\bar{\sigma}(\vec{x}) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}.$$

După cum ştim,

$$M\left(\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}\right) = \sigma^2,$$

$$D^2\left(\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}\right) = \frac{2\sigma^4}{n-1};$$

aşadar, $\bar{\sigma}(\vec{x})$ este o funcţie de estimaţie absolut corectă pentru σ^2.

3. *Metoda verosimilităţii maxime.* Avem

$$\ln P = -\frac{\sum (x_i - m)^2}{2\sigma^2} - n \ln \sigma + c,$$

unde c este o constantă în raport cu m şi σ.
Derivînd în raport cu m şi σ, obţinem respectiv

$$\sum_{i=1}^{m} (x_i - m) = 0,$$

$$\frac{\sum_{i=1}^{m} (x_i - m)^2}{\sigma^2} - n = 0;$$

obţinem astfel următoarele funcţii de estimaţie

$$\bar{m} = \frac{\sum_{i=1}^{n} x_i}{n},$$

$$\bar{\sigma} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - m)^2}{n}}.$$

Am văzut că $\bar{m}$ reprezintă o estimaţie absolut corectă a lui m.

Pe de altă parte, din (XVIII.8,1) și din egalitatea

$$D^2(\bar{\sigma}) = \frac{\sigma^2}{n}\left[n-1-2\left[\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}\right]^2\right] \simeq \frac{\sigma^2}{2n}, \qquad \text{(XVIII.8,2)}$$

rezultă că $\bar{\sigma}$ este o funcție de estimație corectă a lui σ.

Ne propunem să arătăm că funcțiile de estimație $\bar{\alpha}$ și $\bar{\sigma}$ sînt asimptotic eficiente.

Avem

$$M\left(\frac{\partial \ln P}{\partial m}\right) = nM\left(\frac{\partial \ln f}{\partial m}\right)^2 = \frac{n}{\sigma^4},$$

$$M\left(\frac{\partial \ln P}{\partial m} \cdot \frac{\partial \ln P}{\partial \sigma}\right) = nM\left(\frac{\partial \ln f}{\partial m} \cdot \frac{\partial \ln f}{\partial \sigma}\right) = 0,$$

$$M\left(\frac{\partial \ln P}{\partial \sigma}\right)^2 = nM\left(\frac{\partial \ln f}{\partial \sigma}\right)^2 = \frac{2n}{\sigma^2},$$

de unde rezultă că elipsa care dă minimul de concentrație al dispersiei este

$$\frac{n}{\sigma^2}(x-m)^2 + \frac{2n}{\sigma^2}(x-\sigma)^2 = 4. \qquad \text{(XVIII.8,3)}$$

Pe de altă parte, elipsa de concentrație pentru $\overline{m}$ și $\bar{\sigma}$ rezultă din formulele

$$D^2(\overline{m}) = \frac{\sigma^2}{n},$$

$$D^2(\bar{\sigma}) = \frac{\sigma^2}{n}\left[n-1-2\left(\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}\right)^2\right],$$

$$\rho(\overline{m}, \bar{\sigma}) = 0.$$

Ecuația ei este

$$\frac{n}{\sigma^2}(x-m)^2 + \frac{n}{\sigma^2\left[n-1-2\left(\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}\right)^2\right]}(y-\sigma)^2 = 4. \qquad \text{(XVIII.8.4)}$$

Ținînd seamă de (XVIII.8,2), se vede imediat că la limită raportul ariilor celor două elipse tinde către 1.

Calculul este mult mai simplu dacă ne propunem să estimăm perechea de parametri (m, σ^2) în loc de (m, σ) cum am făcut pînă acum.

Ecuația verosimilității maxime ne conduce în același mod la funcțiile de estimație

$$\bar{\alpha} = \frac{\sum_{i=1}^{n} x_i}{n}, \quad \bar{\beta} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}.$$

Avem

$$D^2(\bar{\alpha}) = \frac{\sigma^2}{n}, \quad D^2(\bar{\beta}) = \frac{2\sigma^4}{n-1}, \quad \rho(\bar{\alpha}\bar{\beta}) = 0.$$

Pe de altă parte

$$M\left(\frac{\partial \ln P}{\partial m}\right)^2 = \frac{n}{\sigma^2}, \quad M\left(\frac{\partial \ln P}{\partial m} \cdot \frac{\partial \ln P}{\partial \sigma^2}\right) = 0, \quad M\left(\frac{\partial \ln P}{\partial \sigma^2}\right) = \frac{1}{2\sigma^4}.$$

Elipsa de minimă concentrație a dispersiei are deci drept ecuație

$$\frac{n(x-m)^2}{\sigma^2} + \frac{n(y-\sigma^2)^2}{2\sigma^4} = 4.$$

Elipsa de concentrație a lui $\bar{\alpha}$ și $\bar{\beta}$ are ecuația

$$\frac{(x-m)^2}{\sigma^2} + \frac{n-1}{n} \frac{(y-\sigma^2)^2}{2\sigma^2} = \frac{4}{n}.$$

Cînd n crește necontenit, raportul ariilor celor două elipse tinde către 1. Prin urmare $\bar{\alpha}$ și $\bar{\beta}$ sînt funcții de estimație asimptotic eficiente pentru m și σ^2.

§ 9. Distribuția lui Cauchy

Referitor la distribuția lui Cauchy

$$f(x; \theta) = \frac{1}{\pi} \cdot \frac{1}{1+(x-\theta)^2} \qquad \text{(XVIII.9,1)}$$

ne propunem să estimăm pe θ, în baza a n observații independente, care ne-au condus la valorile

$$x_1, x_2, \ldots, x_n. \qquad \text{(XVIII.9,2)}$$

Media aritmetică a valorilor

$$\bar{x}=\frac{x_1+x_2+\ldots+x_n}{n}$$

nu este o funcție de estimație, deoarece distribuția lui $\bar{x}$ este identică cu (XVIII.9,1). Nu cîștigăm prin urmare în precizie prin creșterea numărului de observații.

Ecuația de verosimilitate maximă este

$$\sum_{i=1}^{n}\frac{(x_i-\theta)}{1+(x_i-\theta)^2}=0.$$

Este foarte greu să obținem o funcție de estimație pentru θ prin rezolvarea acestei ecuații.

Putem însă folosi mediana de selecție ξ_n a valorilor (XVIII.9.2).

Vom vedea la cap. XXII, § 2, că

$$M(\xi_n)=\theta,$$

$$D^2(\xi_n)=\frac{\pi^2}{4n}.$$

Prin urmare, ξ_n este o funcție de estimație absolut corectă. Putem calcula eficiența lui ξ_n.

Intr-adevăr, avem

$$n\int_{-\infty}^{+\infty}\frac{\partial^2 \ln f(x,\theta)}{\partial\theta^2}f(x,\theta)\,d\theta=\frac{n}{\pi}\int_{-\infty}^{+\infty}\frac{2(x-\theta)^2-2}{[1+(x-\theta)^2]^3}\,dx=-\frac{n}{2},$$

de unde rezultă

$$e(\xi_n)=\frac{2}{n}:\frac{\pi^2}{4n}=\frac{8}{\pi^2}=0{,}81.$$

§ 10. Distribuția normală cu două dimensiuni

Să considerăm distribuția normală cu două dimensiuni

$$f(x,y)=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}e^{-\frac{1}{1-\rho^2}\left[\frac{(x-m_1)^2}{\sigma_1^2}-\frac{2\rho(x-m_1)(y-m_2)}{\sigma_1\sigma_2}+\frac{(y-m_2)^2}{\sigma_2^2}\right]},$$

în care există cinci parametri necunoscuți $m_1, m_2, \sigma_1, \sigma_2, \rho$, a căror semnificație o cunoaștem.

Făcînd n experienţe independente, am obţinut valorile experimentale

$$x_1, x_2, \ldots, x_n; \quad y_1, y_2, \ldots, y_n.$$

Ne propunem să găsim, cu ajutorul acestor date, funcţiile de estimaţie pentru evaluarea celor cinci parametri necunoscuţi.
Funcţia de verosimilitate maximă este

$$\ln P = -\frac{1}{1-\rho^2}\left[\frac{1}{\sigma_1^2}\sum(x_i-m_1)^2 - \frac{2\rho}{\sigma_1\sigma_2}\sum(x_i-m_1)(y_i-m_2) + \right.$$

$$\left. + \frac{1}{\sigma_2^2}\sum(y_i-m_2)^2\right] - n\ln\sigma_1 - n\ln\sigma_2 - \frac{n}{2}\ln(1-\rho^2) + A,$$

unde A este o constantă faţă de parametrii consideraţi.
Avem de rezolvat sistemul

$$\frac{\partial \ln P}{\partial m_1}=0, \quad \frac{\partial \ln P}{\partial m_2}=0, \quad \frac{\partial \ln P}{\partial \sigma_1}=0, \quad \frac{\partial \ln P}{\partial \sigma_2}=0, \quad \frac{\partial \ln P}{\partial \rho}=0,$$

care ne dă următoarele funcţii de estimaţie

$$\bar{m}_1 = \frac{\sum x_i}{n} = \bar{x},$$

$$\bar{m}_2 = \frac{\sum y_i}{n} = \bar{y},$$

$$\bar{\sigma}_1 = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2},$$

$$\bar{\sigma}_2 = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2},$$

$$\bar{\rho}\,\bar{\sigma}_1\bar{\sigma}_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}).$$

După teoria generală, aceste funcţii de estimaţie sînt asimptotic eficiente. Verificarea directă a acestei proprietăţi în cazul de faţă o lăsăm pe seama cititorului.

Capitolul XIX

Regiuni de încredere

§ 1. Considerații generale

Să presupunem că avem o familie de distribuții F_α unidimensionale, care depind de parametrul α și a căror funcție de frecvență este $f(x;\alpha)$. In același timp, presupunem că valoarea parametrului α este dată, dar ne este necunoscută și vrem să determinăm valoarea adevărată a parametrului α, printr-o selecție de n probe independente

$$x_1, x_2, \dots, x_n.$$

Procedeul folosit în teoria estimației constă în alegerea unei funcții de selecție $\varphi(x_1, x_2, \dots, x_n)$. Dacă, pentru $n\to\infty$, funcția $\varphi(x_1, x_2, \dots, x_n)$ tinde în probabilitate către α, spunem că $\varphi(x_1, x_2, \dots, x_n)$ este o funcție de estimație pentru α. De aceea funcția $\varphi(x_1, x_2, \dots, x_n)$ poartă și numele de o *statistică a lui* α.

In adevăr, dacă am realizat efectiv experiențele

$$x_1, x_2, \dots, x_n$$

care au luat diferite valori numerice și înlocuim aceste valori numerice în $\varphi(x_1, x_2, \dots, x_n)$, găsim pentru n destul de mare valoarea exactă a parametrului α.

Există însă și un alt mod de a obține indicații asupra mărimii acestui parametru. Să presupunem astfel că efectuînd o selecție de mărime n

$$(x_1, x_2, \dots, x_n),$$

există două funcții de selecție

$$\alpha_1(x_1, x_2, \dots, x_n), \quad \alpha_2(x_1, x_2, \dots, x_n)$$

în așa fel încît probabilitatea inegalității

$$\alpha_1(x_1, x_2, \dots, x_n) \leqq \alpha \leqq \alpha_2(x_1, x_2, \dots x_n) \qquad \text{(XIX.1,1)}$$

să fie independentă de α, adică

$$P[\alpha_1(x_1, x_2, \dots, x_n) \leqq \alpha \leqq \alpha_2(x_1, x_2, \dots, x_n)] = \delta,$$

unde δ nu depinde de α.

Scanned with OKEN Scanner

Dacă acest număr δ este foarte apropiat de 1, înseamnă că inegalitatea (XIX.1,1) este îndeplinită în majoritatea cazurilor. Cum însă α este cuprins între funcţiile α_1 şi α_2, care pentru o selecţie realizată iau valori bine determinate, rezultă că am găsit astfel un interval care acoperă pe α cu o probabilitate δ, apropiată de 1.

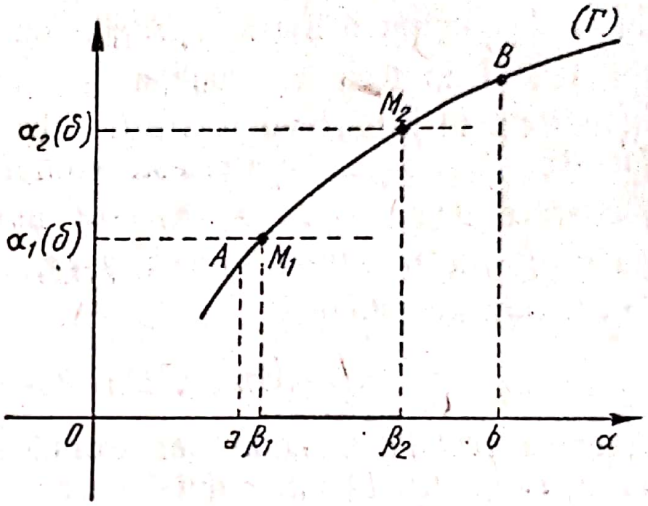


Fig. XIX.1.1

Prin urmare găsim un interval bine determinat în care se află valoarea parametrului α necunoscută nouă. Cu cît acest interval este mai mic şi probabilitatea δ este mai mare, cu atît avem o indicaţie mai precisă cu privire la valoarea lui α. Acest interval se numeşte *interval de încredere.*

Este necesar să atragem atenţia asupra modului cum trebuie înţeleasă legătura dintre valoarea adevărată a lui α şi intervalul de încredere. O afirmaţie de felul următor:

„Probabilitatea ca α să se găsească în intervalul de încredere $[\alpha_1, \alpha_2]$ este egală cu δ" nu este corectă. Parametrul α nu este o variabilă aleatorie, deoarece el are o valoare bine determinată. Faptul că noi nu cunoaştem această valoare nu înseamnă că ea este variabilă. In schimb intervalul $[\alpha_1, \alpha_2]$ este o variabilă aleatorie, fiindcă valorile lui depind de variabilele de selecţie

$$x_1, x_2, \ldots, x_n.$$

De aceea exprimarea justă trebuie să fie astfel

„Probabilitatea ca intervalul $[\alpha_1, \alpha_2]$ să acopere punctul α este egală cu δ".

In aplicaţiile cele mai obişnuite, intervalele de încredere pot fi găsite în următoarele două cazuri.

1. Să admitem că există o funcţie $g(x_1, x_2, \ldots, x_n; \alpha)$ de variabilele selecţiei şi de parametrul α, dar a cărei distribuţie nu depinde de α. Cu alte cuvinte, probabilitatea ca să avem

$$\alpha_1 \leqslant g(x_1, x_2, \ldots, x_n; \alpha) \leqslant \alpha_2$$

este independentă de α. Putem atunci găsi două numere $\alpha_1(\delta)$ şi $\alpha_2(\delta)$ depinzînd de δ, astfel ca

$$P[\alpha_1(\delta) \leq g(x_1, x_2, \dots, x_n; \alpha) \leq \alpha_2(\delta)] = \delta,$$

unde alegem pe δ foarte apropiat de 1.

Să facem ipoteza că funcţia $g(x_1, x_2, \dots, x_n; \alpha)$ este o funcţie monotonă în raport cu α. Pentru fixarea ideilor, să presupunem că funcţia $g(x_1, x_2, \dots, x_n; \alpha)$ este monoton crescătoare în raport cu α. Deoarece $g(x_1, x_2, \dots, x_n; \alpha)$ este monoton crescătoare în raport cu α, există o valoare unică $\beta_1(x_1, x_2, \dots, x_n; \delta)$ astfel ca să fie îndeplinită inegalitatea

$$\alpha_1(\delta) \leq g(x_1, x_2, \dots, x_n; \alpha). \qquad \text{(XIX.1,2)}$$

Aceasta rezultă limpede din examinarea figurii (XIX.1,1). Fie, în acest scop, $[a, b]$ intervalul în care α este definit şi în care funcţia $g(x_1, x_2, \dots, x_n; \alpha)$ este monoton crescătoare în raport cu α. Ducînd o paralelă la axa $O\alpha$ prin punctul $(0, \alpha_1(\delta))$, ea taie curba (Γ), a cărei ecuaţie este $y = g(x_1, x_2, \dots, x_n; \alpha)$ într-un singur punct M_1 de abscisă $\beta_1(x_1, x_2, \dots, x_n; \delta)$. Abscisa β_1 este evident o funcţie de variabilele de selecţie $x_1, x_2, \dots, x_n$ şi de numărul δ. Valorile funcţiei $g(x_1, x_2, \dots, x_n; \alpha)$ mai mari decît $\alpha_1(\delta)$ sînt situate pe arcul de curbă $\overset{\frown}{M_1 B}$, de unde rezultă că, pentru aceste valori, are loc inegalitatea

$$\alpha \geq \beta_1(x_1, x_2, \dots, x_n; \delta). \qquad \text{(XIX.1,3)}$$

Reciproc, numai pentru valorile lui α care satisfac inegalitatea (XIX.1,3) este satisfăcută şi inegalitatea (XIX.1,2). Prin urmare, inegalităţile (XIX.1,2) şi (XIX.1,3) sînt echivalente.

În mod analog, ducînd o paralelă la axa $O\alpha$ prin punctul $(0, \alpha_2(\delta))$, ea taie curba (Γ) în punctul M_2 a cărui abscisă este $\beta_2(x_1, x_2, \dots, x_n; \delta)$, funcţie, de asemenea, de variabilele selecţiei şi numărul δ. Raţionînd la fel ca mai sus, deducem că inegalităţile

$$g(x_1, x_2, \dots, x_n; \alpha) \leq \alpha_2(\delta) \qquad \text{(XIX.1,4)}$$

şi

$$\alpha \leq \beta_2(x_1, x_2, \dots, x_n; \delta) \qquad \text{(XIX.1,5)}$$

sînt echivalente.

Din inegalitățile (XIX.1,2) și (XIX.1,4) pe de o parte și din inegalitățile (XIX.1,3) și (XIX.1,5) echivalente respectiv cu acestea, pe de altă parte, urmează că inegalitățile

$$\alpha_1(\delta) \leqq g(x_1, x_2, \dots, x_n; \alpha) \leqq \alpha_2(\delta)$$

sînt, la rîndul lor, echivalente cu inegalitățile

$$\beta_1(x_1, x_2, \dots, x_n; \delta) \leqq \alpha \leqq \beta_2(x_1, x_2, \dots, x_n; \delta).$$

Atunci, putem scrie

$$P[\beta_1(x_1, x_2, \dots, x_n; \delta) \leqq \alpha \leqq \beta_2(x_1, x_2, \dots, x_n; \delta)] = \delta$$

și deci am determinat un interval de încredere $[\beta_1, \beta_2]$ pentru α.

2. Fie $g(x_1, x_2, \dots, x_n)$ o funcție de selecție a cărei funcție de distribuție este

$$P[g(x_1, x_2, \dots, x_n) \leqq x] = H(x; \alpha).$$

Admitem că există în fiecare punct al distribuției lui $g(x_1, x_2, \dots, x_n)$ o densitate de probabilitate continuă $h(x, \alpha)$.

Fie, de asemenea, r și s două numere pozitive $r \geqq 0$, $s \geqq 0$, $r+s=1$ și $a_1(\alpha; \delta)$ și $b_1(\alpha; \delta)$ două funcții de α și δ pentru care au loc egalitățile

$$\int_{a}^{a_1(\alpha;\,\delta)} h(x; \alpha)\, dx = r(1-\delta) \qquad \text{(XIX.1,6)}$$

și

$$\int_{b_1(\alpha;\,\delta)}^{b} h(x; \alpha)\, dx = s(1-\delta). \qquad \text{(XIX.1,7)}$$

Dacă $[a, b]$ este intervalul de existență al funcției $g(x_1, x_2, \dots, x_n)$ avem evident

$$\int_{a}^{b} h(x; \alpha)\, dx = 1. \qquad \text{(XIX.1,8)}$$

Din egalitățile (XIX.1,6), (XIX.1,7) și (XIX.1,8), deducem

$$\int_{a_1(\alpha;\,\delta)}^{b_1(\alpha;\,\delta)} h(x; \alpha)\, dx = \delta. \qquad \text{(XIX.1,9)}$$

Scanned with OKEN Scanner

Presupunînd că funcţiile $a_1(\alpha;\delta)$ şi $b_1(\alpha;\delta)$ sînt monoton crescătoare în raport cu α, se poate arăta că există un interval de încredere.

In adevăr, în baza egalităţii (XIX.1,9) putem scrie

$$P[a_1(\alpha;\delta) \leqslant g(x_1, x_2, \ldots, x_n) \leqslant b_1(\alpha;\delta)] = \delta,$$

adică probabilitatea inegalităţii

$$a_1(\alpha;\delta) \leqslant g(x_1, x_2, \ldots, x_n) \leqslant b_1(\alpha;\delta) \qquad \text{(XIX.1,10)}$$

este independentă de α.

Funcţia $a_1(\alpha;\delta)$ fiind monoton crescătoare în raport cu α, există, după cum am văzut mai înainte (cazul 1), un număr $A(x_1, x_2, \ldots, x_n;\delta)$ astfel ca inegalităţile

$$\begin{cases} A(x_1, x_2, \ldots, x_n;\delta) \leqslant \alpha, \\ a_1(\alpha;\delta) \leqslant g(x_1, x_2, \ldots, x_n) \end{cases} \qquad \text{(XIX.1,11)}$$

să fie echivalente.

La fel, funcţia $b_1(\alpha;\delta)$ fiind monoton crescătoare în raport cu α, există un număr $B(x_1, x_2, \ldots, x_n;\delta)$ în aşa fel încît inegalităţile

$$\begin{cases} \alpha \leqslant B(x_1, x_2, \ldots, x_n;\delta) \\ g(x_1, x_2, \ldots, x_n \leqslant b_1(\alpha;\delta)) \end{cases} \qquad \text{(XIX.1,12)}$$

sînt echivalente.

Rezultă atunci din (XIX.1,11) şi (XIX.1,12) că inegalitatea (XIX.1,10) este, la rîndul ei, echivalentă cu inegalitatea

$$A(x_1, x_2, \ldots, x_n;\delta) \leqslant \alpha \leqslant B(x_1, x_2, \ldots, x_n;\delta)$$

şi prin urmare

$$P[A(x_1, x_2, \ldots, x_n;\delta) \leqslant \alpha \leqslant B(x_1, x_2, \ldots, x_n;\delta)] = \delta.$$

Am determinat, astfel, intervalul de încredere $[A, B]$, pentru parametrul α.

Este de observat că, în cazul cînd distribuţia teoretică a selecţiei este discontinuă, în loc să considerăm un interval de încredere care acoperă cu probabilitatea δ parametrul α, vom considera un interval de încredere care acoperă parametrul α cu o probabilitate mai

mare sau cel puțin egală cu δ. In acest caz, semnul $=$ (egal) în egalitățile (XIX.1,6) și (XIX.1,7) va fi înlocuit prin semnul $\leq$ (mai mic sau cel mult egal), iar integrala se înlocuiește prin suma obișnuită $(\sum)$.

Dacă în distribuția teoretică sînt mai mulți parametri

$$\alpha_1, \alpha_2 \dots, \alpha_k,$$

astfel ca să avem

$$P[A(x_1, x_2, \dots, x_n; \delta) \leq \alpha_1 \leq B(x_1, x_2, \dots, x_n; \delta)] = \delta,$$

oricare ar fi $\alpha_2, \alpha_3, \dots, \alpha_k$, spunem că $[A, B]$ este un interval de încredere pentru parametrul α_1. Această probabilitate mai poate fi scrisă sub forma

$$P[\alpha_1 \in I(x_1, x_2, \dots, x_n; \delta)] = \delta,$$

notînd prin $I(x_1, x_2, \dots, x_n; \delta)$ intervalul $[A(x_1, x_2, \dots, x_n; \delta), B(x_1, x_2, \dots, x_n; \delta)]$. Parametrii $\alpha_2, \alpha_3, \dots, \alpha_k$ care intervin de prisos în această problemă, îi putem numi *parametri inutili*.

Ca o problemă mai generală, să considerăm $h<k$ parametri

$$\alpha_1, \alpha_2, \dots, \alpha_h$$

dintre cei k parametri

$$\alpha_1, \alpha_2, \dots, \alpha_h, \dots, \alpha_k$$

ai distribuției teoretice.

Fie, în spațiul euclidian cu h dimensiuni R_h, punctul $M(\alpha_1, \alpha_2, \dots; \alpha_h)$ ale cărui coordonate sînt respectiv $\alpha_1, \alpha_2, \dots, \alpha_h$.

In mod analog, dacă în spațiul R_h are loc egalitatea

$$P[M(\alpha_1, \alpha_2, \dots, \alpha_h) \in I(x_1, x_2, \dots, x_n; \delta)] = \delta,$$

oricare ar fi $\alpha_{h+1}, \alpha_{h+2}, \dots, \alpha_k$, spunem că $I(x_1, x_2, \dots, x_n; \delta)$ este o regiune de încredere pentru parametrii $\alpha_1, \alpha_2, \dots, \alpha_h$. Parametrii $\alpha_{h+1}, \alpha_{h+2}, \dots, \alpha_k$, care intervin în mod inutil în distribuția teoretică, sînt ceea ce numim parametri inutili.

In cazul $h=k$, vom obține în spațiul euclidian cu k dimensiuni R_k, o regiune de încredere pentru toți parametrii $\alpha_1, \alpha_2, \dots, \alpha_k$.

Scanned with OKEN Scanner

§ 2. Diferite aplicaţii

1. Să presupunem că distribuţia teoretică este normală $N(m, 1)$ cu funcţia de frecvenţă

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2}},$$

adică o distribuţie cu un singur parametru m.

Să considerăm funcţia de selecţie

$$\frac{\bar{x}-m}{\sqrt{n}}$$

care, după cum se constată uşor, corespunde unei distribuţii normale $N\left(0, \frac{1}{n}\right)$. Prin urmare, distribuţia acestei funcţii de selecţie nu depinde de parametrul m şi, pe lîngă aceasta, este şi monoton descrescătoare în rapot cu m.

Sîntem astfel în condiţiile cazului 1 din paragraful precedent şi putem, atunci, să construim un interval de încredere pentru parametrul m. Avem, deci, egalitatea

$$P\left[\alpha \leqslant \frac{\bar{x}-m}{\sqrt{n}} \leqslant \beta\right] = \frac{n}{\sqrt{2\pi}} \int_{\alpha}^{\beta} e^{-\frac{n^2x^2}{2}} dx.$$

Putem să ne alegem numerele α şi β în aşa fel încît să aibă loc egalitatea

$$\frac{n}{\sqrt{2\pi}} \int_{\alpha}^{\beta} e^{-\frac{n^2x^2}{2}} dx = \delta, \qquad \text{(XIX.2,1)}$$

δ fiind un număr foarte apropiat de 1, de unde rezultă că

$$P\left(\alpha \leqslant \frac{\bar{x}-m}{\sqrt{n}} \leqslant \beta\right) = \delta,$$

care se mai poate pune sub forma

$$P(\bar{x}-\beta\sqrt{n} \leqslant m \leqslant \bar{x}-\alpha\sqrt{n}) = \delta.$$

Am construit, aşadar, intervalul de încredere

$$[\bar{x}-\beta\sqrt{n},\ \bar{x}-\alpha\sqrt{n}]$$

pentru parametrul m.

Este de observat că se pot obţine o infinitate de intervale de încredere pentru parametrul m, toate cu aceeaşi probabilitate δ, deoarece numerele α şi β verifică o singură ecuaţie dată de (XIX.2,1). Este natural, însă, să acordăm o preferinţă intervalului cel mai convenabil.

Pentru alegerea intervalului de încredere cel mai convenabil, există şi alte criterii asimptotice, ca acelea pe care le vom examina în teoria testelor de verificare.

2. Să presupunem acum că distribuţia teoretică este normală $N(0, \sigma)$ cu funcţia de frecvenţă

$$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}},$$

adică o distribuţie teoretică de asemenea cu un singur parametru σ. In acest caz, să considerăm funcţia de selecţie

$$\frac{x_1^2+x_2^2+\ldots+x_n^2}{\sigma^2}$$

a cărei distribuţie corespunde unei legi $H(n, 1)$. In adevăr, observăm că funcţia de selecţie este o sumă de n pătrate $\frac{x^2}{\sigma^2}$, unde fiecare variabilă de selecţie $\frac{x_i}{\sigma}$ $(i=1, 2, \ldots, n)$ este distribuită normal $N(0, 1)$. Prin urmare, funcţia de selecţie considerată admite o distribuţie independentă de σ şi, în acelaşi timp, este monoton descrescătoare în raport cu σ^2.

Atunci, fiind îndeplinite condiţiile primului caz din paragraful precedent, putem, evident, determina un interval de încredere pentru parametrul σ.

Avem, astfel, egalitatea

$$P\left[\alpha \leqslant \frac{x_1^2+x_2^2+\ldots+x_n^2}{\sigma^2} \leqslant \beta\right] = \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} \int_{\alpha}^{\beta} x^{\frac{n}{2}-1} e^{-\frac{x}{2}} dx.$$

Numerele α şi β le alegem în aşa fel încît să avem îndeplinită egalitatea

$$\frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)}\int_{\alpha}^{\beta} x^{\frac{n}{2}-1}e^{-\frac{x}{2}}dx=\delta, \qquad \text{(XIX.2,2)}$$

unde δ este foarte aproape de 1.

Rezultă deci că

$$P\left(\alpha\leqslant\frac{x_1^2+x_2^2+\ldots+x_n^2}{\sigma^2}\leqslant\beta\right)=\delta,$$

sau, sub altă formă,

$$P\left(\sqrt{\frac{\sum_{i=1}^{n}x_i^2}{\beta}}\leqslant\sigma\leqslant\sqrt{\frac{\sum_{i=1}^{n}x_i^2}{\alpha}}\right)=\delta.$$

In acest mod, am determinat o infinitate de intervale de încredere pentru parametrul σ, deoarece, folosind aceeaşi observaţie de la punctul 1, numerele α şi β verifică o singură ecuaţie dată de (XIX.2,2).

3. Presupunem, în sfîrşit, că distribuţia teoretică este normală $N(m, \sigma)$ cu funcţia de frecvenţă

$$\frac{1}{\sqrt{2\pi}\sigma}\cdot e^{-\frac{(x-m)^2}{2\sigma^2}},$$

adică o distribuţie teoretică cu doi parametri m şi σ. Ne propunem să determinăm cîte un interval de încredere pentru fiecare dintre parametrii m şi σ.

Incepem cu determinarea intervalului de încredere pentru parametrul m. Metoda întrebuinţată la punctul 1, pentru găsirea intervalului de încredere al parametrului m, nu mai este indicată aici, deoarece nu cunoaştem valoarea lui σ. De aceea, vom considera funcţia de selecţie

$$\sqrt{n}\,\frac{\bar{x}-m}{\sqrt{m_2}}, \qquad \text{(XIX.2,3)}$$

care, în baza celor arătate la studiul distribuţiei lui Student (cap. XI, § 9), urmează distribuţia lui Student cu $n-1$ grade de libertate, avînd funcţia de frecvenţă

$$\frac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{\pi(n-1)}\,\Gamma\left(\frac{n-1}{2}\right)}\left(1+\frac{t^2}{n-1}\right)^{-\frac{n}{2}}.$$

Pe de o parte, funcţia de selecţie (XIX.2,3) are o distribuţie independentă de σ, iar pe de altă parte este o funcţie monoton descrescătoare în raport cu m. Prin urmare, fiind îndeplinite condiţiile cazului întîi din paragraful precedent, putem scrie

$$P\left(\alpha \leqslant \sqrt{n}\,\frac{\bar{x}-m}{\sqrt{m_2}} \leqslant \beta\right) = \frac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{\pi(n-1)}\,\Gamma\left(\frac{n-1}{2}\right)}\int_{\alpha}^{\beta}\left(1+\frac{t^2}{n-1}\right)^{-\frac{n}{2}} dt.$$

Luînd numerele α şi β astfel ca

$$\frac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{\pi(n-1)}\,\Gamma\left(\frac{n-1}{2}\right)}\int_{\alpha}^{\beta}\left(1+\frac{t^2}{n-1}\right)^{-\frac{n}{2}} dt=\delta \qquad (\delta \simeq 1)$$

urmează că

$$P\left(\alpha \leqslant \sqrt{n}\,\frac{\bar{x}-m}{m_2} \leqslant \beta\right)=\delta$$

sau, ceea ce este tot una, cu

$$P\left(\bar{x}-\frac{\beta}{\sqrt{n}}\sqrt{m_2} \leqslant m \leqslant \bar{x}-\frac{\alpha}{\sqrt{n}}\sqrt{m_2}\right)=\delta.$$

Se vede că, şi în acest caz, putem obţine o infinitate de intervale de încredere pentru parametrul m. Aceste intervale de încredere sînt determinate, oricare ar fi valoarea parametrului σ, care este deci un parametru inutil. Se observă că, dacă n creşte, intervalele de încredere

$$\left[\bar{x}-\frac{\beta}{\sqrt{n}}\sqrt{m_2},\quad \bar{x}-\frac{\alpha}{\sqrt{n}}\sqrt{m_2}\right]$$

devin din ce în ce mai mici.

Să determinăm acum un interval de încredere pentru parametrul σ. In acest scop, să considerăm statistica

$$(x_1-\bar{x})^2+(x_2-\bar{x})^2+\ldots+(x_n-\bar{x})^2,$$

care, potrivit celor arătate la studiul selecţiei dintr-o colectivitate normală (cap. XIII), corespunde unei distribuţii $H(n-1, \sigma)$. Atunci statistica sau, altfel, funcţia de selecţie

$$\frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\ldots+(x_n-\bar{x})^2}{\sigma^2}$$

corespunde, evident, unei distribuţii $H(n-1, 1)$. După cum se observă, această funcţie de selecţie este monoton descrescătoare în raport cu σ^2, iar distribuţia ei este independentă de σ, ceea ce ne permite să scriem că

$$P\left[\alpha \leqslant \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{\sigma^2} \leqslant \beta\right] = \frac{1}{2^{\frac{n-1}{2}}\Gamma\left(\frac{n-1}{2}\right)}\int_{\alpha}^{\beta} x^{\frac{n-3}{2}} e^{-\frac{x^2}{2}}dx,$$

numerele α şi β verificînd ecuaţia

$$\frac{1}{2^{\frac{n-1}{2}}\Gamma\left(\frac{n-1}{2}\right)}\int_{\alpha}^{\beta} x^{\frac{n-3}{2}} e^{-\frac{x^2}{2}}dx=\delta \qquad (\delta \simeq 1).$$

Urmează deci că

$$P\left(\alpha \leqslant \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{\sigma^2} \leqslant \beta\right)=\delta$$

sau

$$P\left(\sqrt{\frac{1}{\beta}\sum_{i=1}^{n}(x_i-\bar{x})^2} \leqslant \sigma \leqslant \sqrt{\frac{1}{\alpha}\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)=\delta.$$

Prin urmare, obţinem o infinitate de intervale de încredere şi pentru parametrul σ, care se pot determina oricare ar fi valoarea lui m. Parametrul m, în acest caz, este un parametru inutil.

4. De data aceasta, să considerăm cazul cînd distribuţia teoretică este discontinuă. Fie, astfel, distribuţia binomială, unde probabili-

tatea ca în n probe independente să obţinem valoarea α, conform celor arătate la (cap. VII, § 2), este egală cu

$$C_n^\alpha p^\alpha q^{n-\alpha} \qquad \left(\sum_{\alpha=0}^{n} C_n^\alpha p^\alpha q^{n-\alpha} = 1\right).$$

Probabilitatea p de a se verifica în fiecare probă evenimentul E ne este necunoscută şi ne propunem să o estimăm, construind un interval de încredere. Pentru determinarea acestui interval de încredere, vom aplica metoda expusă la cazul al doilea din paragraful precedent.

Să considerăm inegalităţile

$$\sum_{\alpha=0}^{a_1(p,\,\delta)} C_n^\alpha p^\alpha q^{n-\alpha} \leqslant r(1-\delta)$$

şi

$$\sum_{\alpha=b_1(p,\,\delta)}^{n} C_n^\alpha p^\alpha q^{n-\alpha} \leqslant s(1-\delta),$$

unde r şi s sînt două numere pozitive $(r \geqslant 0)$, $(s \geqslant 0)$ şi $(r+s=1)$, sumele înţelegîndu-se pînă la (respectiv de la) cel mai mare număr cuprins în $a_1(p, \delta)$ şi $b_1(p, \delta)$.

Deducem că

$$\sum_{\alpha=a_1(p,\,\delta)}^{b_1(p,\,\delta)} C_n^\alpha p^\alpha q^{n-\alpha} \geqslant \delta.$$

Presupunînd că $p' > p$, avem

$$\sum_{\alpha=0}^{a_1(p',\,\delta)} C_n^\alpha p'^\alpha q'^{n-\alpha} \leqslant r(1-\delta)$$

şi prin urmare

$$\sum_{\alpha=0}^{a_1(p,\,\delta)} C_n^\alpha p^\alpha q^{n-\alpha} = \sum_{\alpha=0}^{a_1(p',\,\delta)} C_n^\alpha p'^\alpha q'^{n-\alpha}. \qquad \text{(XIX.2,4)}$$

Deoarece funcţia de distribuţie

$$F(k) = \sum_{\alpha=0}^{k} C_n^\alpha p^\alpha q^{n-\alpha} = 1 - \frac{\int_0^p x^k (1-x)^{n-k-1} dx}{\int_0^1 x^k (1-x)^{n-k-1} dx}$$

este o funcţie monoton descrescătoare în raport cu p, rezultă că

$$\sum_{a=0}^{a_1(p,\,\delta)} C_n^a p^a q^{n-a} > \sum_{a=0}^{a_1(p,\,\delta)} C_n^a p'^a q'^{n-a}$$

sau, ţinînd seamă de egalitatea (XIX.2,4),

$$\sum_{a=0}^{a_1(p',\,\delta)} C_n^a p'^a q'^{n-a} > \sum_{a=0}^{a_1(p,\,\delta)} C_n^a p'^a q'^{n-a}.$$

Urmează, deci, că

$$a_1(p',\,\delta) > a_1(p,\,\delta),$$

adică funcţia $a_1(p,\,\delta)$ este o funcţie monoton crescătoare.

In mod analog, se demonstrează că şi funcţia $b_1(p,\,\delta)$ este o funcţie monoton crescătoare.

Atunci, fiind îndeplinite condiţiile arătate la cazul 2 din paragraful precedent, se poate construi un interval de încredere pentru probabilitatea p prin metoda expusă acolo. Cu cît acest interval va fi mai mic şi numărul δ mai apropiat de 1, cu atît vom obţine o indicaţie mai precisă asupra probabilităţii p.

5. Pînă acum am dat exemple numai de intervale de încredere. Vom da acum un exemplu de regiune de încredere, furnizat de diferenţa mediilor a două distribuţii normale.

Să considerăm două selecţii de variabile independente

$$x_1,\ x_2,\ \ldots,\ x_{n_1},$$

$$y_1,\ y_2,\ \ldots,\ y_{n_1},$$

efectuate respectiv din populaţiile normale $N(m_1,\ \sigma_1)$ şi $N(m_2,\ \sigma_2)$. Ne propunem să determinăm o regiune de încredere pentru m_1 şi m_2 în spaţiul parametrilor $(m_1,\ m_2,\ \sigma_1,\ \sigma_2)$.

Notăm

$$\bar{x} = \frac{\sum_{i=1}^{n_1} x_i}{n_1}, \quad \bar{y} = \frac{\sum_{i=1}^{n_2} y_i}{n_2},$$

$$m_2^{(1)} = \frac{\sum_{i=1}^{n_1} (x_i - \bar{x})^2}{n_1}, \quad m_2^{(2)} = \frac{\sum_{i=1}^{n_2} (y_i - \bar{y})^2}{n_2}.$$

Variabilele

$$\sqrt{n_1-1}\,\frac{\bar{x}-m_1}{\sqrt{m_2^{(1)}}},\quad \sqrt{n_2-1}\,\frac{\bar{y}-m_2}{\sqrt{m_2^{(2)}}} \tag{XIX.2,3}$$

sînt distribuite după o lege Student cu respectiv n_1-1 şi n_2-1 grade de libertate.

Să considerăm inegalităţile

$$k_1 \leqslant a_1\sqrt{n_1-1}\,\frac{\bar{x}-m_1}{\sqrt{m_2^{(1)}}} + a_2\sqrt{n_2-1}\,\frac{\bar{y}-m_2}{\sqrt{m_2^{(2)}}} \leqslant k_2, \tag{XIX.2,4}$$

unde a_1, a_2, k_1, k_2 sînt patru constante oarecare, supuse la condiţia $k_1 < k_2$.

Dacă notăm

$$u=\sqrt{n_1-1}\,\frac{\bar{x}-m_1}{\sqrt{m_2^{(1)}}},\quad v=\sqrt{n_2-1}\,\frac{\bar{y}-m_2}{\sqrt{m_2^{(2)}}},$$

probabilitatea ca punctele $(x_1, x_2, \ldots, x_n; y_1, y_2, \ldots, y_n)$ din spaţiul total al selecţiei să satisfacă inegalităţile (XIX.2,4) este

$$\frac{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)}{\pi\sqrt{(n_1-1)(n_2-1)}\,\Gamma\left(\frac{n_1-1}{2}\right)\Gamma\left(\frac{n_2-1}{2}\right)} \iint \left(1+\frac{u^2}{n_1-1}\right)^{-\frac{n_1}{2}} \left(1+\frac{v^2}{n_2-1}\right)^{-\frac{n_2}{2}} du\,dv \tag{XIX.2,5}$$

integrala fiind extinsă la toate punctele din planul (u, v) pentru care avem

$$k_1 \leqslant a_1u+a_2v \leqslant k_2.$$

Inegalitatea (XIX.2,4) se poate scrie sub forma

$$\frac{a_1\bar{x}}{\sqrt{m_1^{(2)}}}\sqrt{n_1-1}+\frac{a_2\bar{y}}{\sqrt{m_2^{(2)}}}\sqrt{n_2-1}-k \leqslant \frac{a_1m_1}{\sqrt{m_1^{(2)}}}\sqrt{n_1-1}+$$

$$+\frac{a_2m_2}{\sqrt{m_2^{(2)}}}\sqrt{n_2-1} \leqslant \frac{a_1\bar{x}}{\sqrt{m_1^{(2)}}}\sqrt{n_1-1}+\frac{a_2\bar{y}}{\sqrt{m_2^{(2)}}}\sqrt{n_2-1}. \tag{XIX.2,6}$$

Pentru orice punct $\bar{x}$ din spaţiul de selecţie, relaţiile de mai sus definesc o regiune din spaţiul parametrilor, pentru care probabilitatea ca punctul (m_1, m_2) să cadă în ea este dată de (XIX.2,5).

Se constată uşor că mulţimea de puncte definită în spaţiul parametrilor de inegalităţile (XIX.2,6) este o regiune de încredere.

Un caz frecvent întîlnit în aplicaţii este acela în care cele două distribuţii teoretice normale au aceeaşi dispersie. Avem prin urmare

$$\sigma_1 = \sigma_2 = \sigma.$$

Variabilele

$$\frac{\sqrt{n_1}\,(\bar{x}-m_1)}{\sigma}, \quad \frac{\sqrt{n_2}\,(\bar{y}-m_2)}{\sigma}$$

urmează respectiv distribuţiile normale $N(0, 1)$.

De aici rezultă că variabilele

$$\frac{n_1\,(\bar{x}-m_1)^2}{\sigma^2}, \quad \frac{n_2\,(\bar{y}-m_2)^2}{\sigma^2} \qquad \text{(XIX.2,7)}$$

sînt distribuite după o lege Helmert-Pearson cu un singur grad de libertate.

Pe de altă parte, variabilele

$$\frac{\overline{m}_2^{(1)}}{\sigma^2} = \frac{\sum_{i=1}^{n_1}(\bar{x}-x_i)^2}{n_1\,\sigma^2}, \quad \frac{\overline{m}_2^{(2)}}{\sigma^2} = \frac{\sum_{i=1}^{n_2}(\bar{y}-y_i)^2}{n_2\,\sigma^2} \qquad \text{(XIX.2,8)}$$

sînt distribuite respectiv după aceleaşi legi $H(n_1-1, \sigma)$, $H(n_2-1, \sigma)$, însă cu n_1-1 şi n_2-1 grade de libertate.

Rezultă atunci că variabilele

$$\frac{n_1\,(\bar{x}-m_1)^2}{\sigma^2} + \frac{n_2\,(\bar{y}-m_2)^2}{\sigma^2},$$

$$\frac{\overline{m}_2^{(1)}}{\sigma^2} + \frac{\overline{m}_2^{(2)}}{\sigma^2}$$

urmează respectiv legile $H(2, 1)$ şi $H(n_1+n_2-2, 1)$.

Pe de altă parte, am arătat că aceste variabile sînt independente între ele.

Prin urmare, variabila

$$\frac{n_1\,(\bar{x}-m_1)^2+n_2\,(y-m_2)^2}{\overline{m}_2^{(1)}+\overline{m}_2^{(2)}} \cdot \frac{n_1+n_2-2}{2}$$

este distribuită după legea Snedecor, cu 2 și n_1+n_2-2 grade de libertate, a cărei funcție de frecvență este

$$\left(\frac{2}{n-2}\right)^{\frac{1}{2}} \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n}{2}-1\right)} \cdot u^{-\frac{1}{2}}\left(1+\frac{2}{n-2}u\right)^{-\frac{n}{2}} \qquad (u \geqq 0),$$

unde am notat $n_1+n_2=n$.

Fixîndu-l pe δ cît vrem de aproape de 1, putem determina numerele (α, β) astfel ca

$$\left(\frac{2}{n-2}\right)^{\frac{1}{2}} \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n}{2}-1\right)} \int_{\alpha}^{\beta} u^{-\frac{1}{2}}\left(1+\frac{2}{n-2}u\right)^{-\frac{n}{2}} du=\delta.$$

Obținem astfel

$$\alpha \frac{2\left[\overline{m}_2^{(1)}+\overline{m}_2^{(2)}\right]}{n-2} \leqq n_1(\overline{x}-m_1)^2+n_2(\overline{y}-m_2)^2 \leqq \beta \frac{2\left[\overline{m}_2^{(1)}+\overline{m}_2^{(2)}\right]}{n-2},$$

care în spațiul parametrilor (m_1, m_2, σ) reprezintă o regiune de încredere pentru m_1 și m_2, cuprinsă între doi cilindri eliptici; σ este un parametru inutil.

§ 3. Construirea intervalelor de încredere pentru selecții mari

Intervalele de încredere, de care ne-am ocupat în paragrafele precedente, au fost determinate pentru orice selecție. Se pot însă determina intervale de încredere numai pentru selecții mari. Vom dovedi, în condiții destul de largi, folosind funcția de verosimilitate maximă, că putem găsi un interval de încredere pentru selecțiile mari.

Fie $f(x;\ \alpha)$ funcția de frecvență a distribuției teoretice cu un singur parametru α și

$$L=f(x_1;\ \alpha)\, f(x_2;\ \alpha) \ldots f(x_n;\ \alpha)$$

funcția de verosimilitate maximă, unde $x_1, x_2, \ldots, x_n$ sînt variabilele de selecție.

Rezultă atunci că

$$\frac{\partial \ln L}{\partial \alpha} = \sum_{i=1}^{n} \frac{\partial \ln f(x_i;\ \alpha)}{\partial \alpha} = \sum_{i=1}^{n} y_i,$$

însemnînd

$$y_i = \frac{\partial \ln f(x_i;\ \alpha)}{\partial \alpha}.$$

Dacă

$$y = \frac{\partial \ln f(x;\ \alpha)}{\partial \alpha}$$

este o variabilă aleatorie, avem

$$M(y) = \int_{-\infty}^{+\infty} \frac{\partial \ln f(x;\ \alpha)}{\partial \alpha} f(x;\ \alpha)\, dx = \int_{-\infty}^{+\infty} \frac{\partial f(x;\ \alpha)}{\partial \alpha} dx = 0. \qquad \text{(XIX.3,1)}$$

In adevăr, din relaţia

$$\int_{-\infty}^{+\infty} f(x;\ \alpha)\, dx = 1,$$

deducem

$$\int_{-\infty}^{+\infty} \frac{\partial f(x;\ \alpha)}{\partial \alpha} dx = 0,$$

presupunînd că este posibilă derivarea sub semnul integrală.
Presupunem, pe de altă parte, că dispersia variabilei aleatorii y

$$D^2(y) = \int_{-\infty}^{+\infty} \left[\frac{\partial \ln f(x;\ \alpha)}{\partial \alpha}\right]^2 \cdot f(x;\ \alpha)\, dx \qquad \text{(XIX.3,2)}$$

este finită şi diferită de zero.
Prin urmare, avînd în vedere relaţiile (XIX.3,1) şi (XIX.3,2), obţinem

$$M\left(\sum_{i=1}^{n} y_i\right) = \sum_{i=1}^{n} M(y_i) = 0$$

şi

$$M\left[\left(\sum_{i=1}^{n} y_i\right)^2\right] = \sum_{i=1}^{n} M(y_i^2) + 2\sum_{i \neq j} M(y_i)\, M(y_j) = nD^2.$$

Atunci, în baza teoremei lui Liapunov, putem să scriem

$$\lim_{n\to\infty} P\left(-a \leqslant \frac{y_1+y_2+\ldots+y_n}{D\sqrt{n}} \leqslant a\right) \to \frac{1}{\sqrt{2\pi}} \int_{-a}^{+a} e^{-\frac{x^2}{2}} dx.$$

Cu alte cuvinte, pentru valorile mari ale lui n, putem afirma că statistica

$$\frac{y_1+y_2+\ldots+y_n}{D\sqrt{n}} = \frac{1}{D\sqrt{n}} \cdot \frac{\partial \ln L}{\partial \alpha} \qquad \text{(XIX.3,3)}$$

urmează o lege normală $N(0, 1)$, dacă D este finit şi diferit de zero.

Dacă, în afară de aceasta, mai admitem că statistica

$$\frac{y_1+y_2+\ldots+y_n}{D\sqrt{n}}$$

este monoton crescătoare, putem astfel construi un interval de încredere pentru parametrul α, după metoda indicată mai înainte la cazul 1 din § 1 al capitolului de faţă.

Să aplicăm modul de construire a intervalelor de încredere pentru selecţii mari, la următoarele două exemple.

1. Să considerăm distribuţia binomială în care nu cunoaştem parametrul p şi pentru care urmează, deci, să construim un interval de încredere cu ajutorul unei selecţii mari.

In acest caz, funcţia de verosimilitate maximă este

$$L = p^{\alpha} \cdot q^{n-\alpha}, \qquad (q = 1-p),$$

unde α reprezintă numărul de cîte ori s-a verificat evenimentul E în cele n probe independente

$$x_1, x_2, \ldots, x_n,$$

de unde rezultă că

$$\frac{\partial \ln L}{\partial p} = \frac{\alpha}{p} - \frac{n-\alpha}{q} = \frac{\alpha - np}{pq}.$$

Cum, însă, fiecare variabilă a selecţiei corespunde unei distribuţii

$$X_i\begin{pmatrix} 1 & 0 \\ p & q \end{pmatrix} \qquad (i=1,\ 2,\ \ldots,\ n),$$

avem evident

$$y_1=\frac{\partial \ln p}{\partial p}=\frac{1}{p},$$

$$y_2=\frac{\partial \ln q}{\partial p}=-\frac{1}{q}$$

şi prin urmare

$$D^2=\frac{1}{p^2}\cdot p+\frac{1}{q^2}\cdot q=\frac{1}{pq}.$$

Atunci, statistica (XIX.3,3) devine

$$\frac{1}{D\sqrt{n}}\cdot\frac{\partial \ln L}{\partial p}=\frac{\alpha-np}{\sqrt{npq}}=\frac{\sqrt{n}\,(\bar{x}-p)}{\sqrt{p\,(1-p)}},$$

ţinînd seama că $\alpha=x_1+x_2+\ldots+x_n$.
Putem astfel să scriem

$$\underset{n\to\infty}{P}\left(-a\leqslant\sqrt{n}\,\frac{\bar{x}-p}{\sqrt{p\,(1-p)}}\leqslant a\right)\to\frac{1}{\sqrt{2\pi}}\int_{-a}^{+a}e^{-\frac{x^2}{2}}\,dx$$

sau sub o altă formă

$$P\,(p_1<p<p_2)\to\frac{1}{\sqrt{2\pi}}\int_{-a}^{+a}e^{-\frac{x^2}{2}}\,dx,$$

unde p_1 şi p_2 sînt rădăcinile ecuaţiei de gradul al doilea în p

$$(n+a^2)p^2-(2n\bar{x}+a^2)\,p+n\bar{x}^2=0.$$

Am găsit, deci pentru parametrul p, cînd $n\to\infty$, intervalul de încredere $[p_1,\ p_2]$ care, după cum se constată uşor din ecuaţia de mai sus, este întotdeauna cuprins în intervalul $[0,\ 1]$.

Scanned with OKEN Scanner

2. Să considerăm, acum, distribuţia Poisson cu funcţia de frecvenţă

$$f(x;\lambda)=\frac{\lambda x e^{-\lambda}}{x!} \qquad (x=0,\ 1,\ldots,\infty).$$

Se cere să se determine un interval de încredere pentru parametrul necunoscut λ.

Atunci, variabila aleatorie y este

$$\frac{\partial \ln f(x;\lambda)}{\partial\lambda}=\frac{x}{\lambda}-1$$

şi deci

$$D^2=\sum_{x=0}^{\infty}\left(\frac{x}{\lambda}-1\right)^2\frac{\lambda^x e^{-\lambda}}{x!}=\frac{1}{\lambda},$$

avînd în vedere valoarea momentelor de ordinul întîi şi deci ale distribuţiei Poisson cu parametrul λ.

In acest caz, expresia statisticii (XIX.3,3) devine

$$\frac{y_1+y_2+\ldots+y_n}{D\sqrt{n}}=\frac{n\left(\frac{\bar{x}}{\lambda}-1\right)\sqrt{\lambda}}{\sqrt{n}}=\sqrt{n}\,\frac{(\bar{x}-\lambda)}{\sqrt{\lambda}},$$

de unde rezultă

$$\underset{n\to\infty}{P}\left(-a\leqslant\sqrt{n}\,\frac{\bar{x}-\lambda}{\sqrt{\lambda}}\leqslant a\right)\to\frac{1}{\sqrt{2\pi}}\int_{-a}^{+a}e^{-\frac{x^2}{2}}dx,$$

ceea ce este echivalent cu

$$P(\lambda_1\leqslant\lambda\leqslant\lambda_2)\to\frac{1}{\sqrt{2\pi}}\int_{-a}^{+a}e^{-\frac{x^2}{2}}dx,$$

notînd prin λ_1 şi λ_2 rădăcinile ecuaţiei de gradul al doilea în raport cu λ

$$n\lambda^2-(2n\bar{x}+a^2)\lambda+n\bar{x}^2=0.$$

Am determinat, astfel, pentru valorile mari ale lui n, intervalul de încredere $[\lambda_1,\ \lambda_2]$ pentru parametrul λ.

Scanned with OKEN Scanner

Capitolul XX

Teoria estimației pentru variabile dependente

§ 1. Considerații generale

Teoria estimației a fost construită pe baza principiului că variabilele $X_i (i=1, 2, \ldots, n)$, care fac parte din selecție, sînt independente între ele. Această restricție a permis să se obțină teoremele atît de simple și elegante, pe care le întîlnim în această ramură a statisticii matematice. Trebuie să se observe, pe de altă parte, că restricția de mai sus este impusă și de considerații de ordin practic, deoarece în majoritatea cazurilor selecțiile întîlnite în statistica practică sînt formate din variabile independente între ele.

In anumite cazuri, însă, realizarea independenței observațiilor care compun o selecție este imposibilă. Să considerăm astfel cazul clasic al măsurării repetate a unei aceleiași mărimi, luată ca exemplu în teoria erorilor. Dacă adevărata mărime este l, iar în n măsurători independente am obținut valorile $l_i (i=1, 2, \ldots, n)$, înseamnă că am făcut erorile $l_i - l (i=1, 2, \ldots, n)$. Este natural să admitem independența observațiilor, care ne-au condus la valorile l_i, pentru că fiecare măsurătoare am făcut-o în mod separat. Totuși, se observă în multe cazuri că două erori consecutive $l_i - l$ și $l_{i+1} - l$ au tendința să fie de același semn, ceea ce s-ar putea explica prin faptul că, în general, condițiile în care se fac două măsurători consecutive sînt mai apropiate decît condițiile referitoare la două măsurători mai depărtate între ele. Prin urmare, chiar în acest caz simplu, de multe ori realizarea independenței observațiilor este imposibilă.

Sînt însă cazuri în care chiar natura fenomenului nu permite realizarea unei selecții independente. Să considerăm astfel o populație microbiană care se dezvoltă după legile contagiunii. Vom observa această populație la datele $t_1, t_2, \ldots, t_m$ și vom obține astfel o selecție. Este evident că datele astfel obținute sînt dependente și că o selecție cu observații independente nu putem realiza. Procedeul constînd în a considera n populații microbiene distincte și a le observa în mod separat conduce desigur la o selecție independentă, dar de multe ori nu avem la îndemînă mai multe populații sau nu sîntem siguri că putem realiza experiențele în astfel de condiții de exactitate, încît populațiile să aibă aceeași distribuție

statistică. Prin urmare, mijlocul cel mai corect de cercetare este de a trage concluzii din selecția care se obține în mod natural, observînd populația la anumite date în timp, adică să considerăm cazul cînd selecția este formată din variabile dependente după legea de probabilitate a fenomenului considerat.

Teoria statisticii pentru variabilele dependente este de-abia la începutul ei. In cele ce urmează, vom expune cîteva rezultate privind estimarea parametrilor, care intră într-un lanț Markov discontinuu sau continuu. Expunerea sumară a teoriei lanțurilor lui Markov, care urmează, ne permite să stabilim aceste rezultate.

§ 2. Lanțuri cu legături complete

Fie

$$X_1, \quad X_2, \ldots, X_n, \ldots \tag{XX.2,1}$$

un șir de variabile dependente.

Aceste variabile se succed în ordinea crescătoare a indicilor. Astfel, rîndul lui X_{n+1} de a fi supus experienței, care decide modul de realizare al fenomenului aleator pe care-l reprezintă această variabilă, nu vine decît după ce toate experiențele referitoare la variabilele $X_1, X_2, \ldots, X_n$ s-au realizat.

Pe de altă parte, să presupunem că variabila X_n reprezintă, oricare ar fi n, același fenomen aleator.

Dacă amîndouă condițiile de mai sus sînt îndeplinite, șirul (XX.2,1) formează, prin definiție, un *lanț*.

Presupunem că variabilele X_n iau valorile

$$a_1, \ a_2, \ldots, a_s.$$

In cazul general, probabilitatea ca X_n să ia valoarea a_{j_n} depinde de valorile

$$a_{j_1}, \ a_{j_2}, \ldots, a_{j_{n-1}}$$

luate de variabilele precedente $X_1, X_2, \ldots, X_{n-1}$.

Probabilitatea ca să avem

$$X_n = a_{j_n},$$

după ce am avut îndeplinite egalitățile

$$X_1 = a_{j_1}, \quad X_2 = a_{j_2}, \ldots, X_{n-1} = a_{j_{n-1}},$$

poartă numele de *probabilitate fundamentală*. In cele ce urmează, o vom nota prin

$$P_{j_1, j_2, \ldots, j_{n-1}; j_n}.$$

Reiese, în baza principiului probabilităţilor totale, că

$$\sum_{j_n=1}^{s} P_{j_1, j_2, \ldots, j_{n-1}; j_n} = 1. \qquad \text{(XX.2,2)}$$

Numărul probabilităţilor fundamentale referitoare la variabila X_n se ridică la s^n, deoarece fiecare dintre indicii $j_1, j_2, \ldots, j_n$ poate varia între 1 şi s inclusiv. Este însă de observat că aceste probabilităţi nu sînt independente între ele, deoarece trebuie să îndeplinească relaţiile (XX.2,2).

Numărul relaţiilor (XX.2,2) se determină însă imediat. El este egal cu s^{n-1}, deoarece pentru fiecare combinaţie distinctă $j_1, j_2, \ldots, j_{n-1}$ se obţine o astfel de relaţie.

Prin urmare, numărul probabilităţilor fundamentale independente ataşate variabilei X_n este egal cu $s^n - s^{n-1} = s^{n-1}(s-1)$.

Putem, deci, să dăm valori arbitrare, bineînţeles cuprinse între 0 şi 1, celor s^{n-1} probabilităţi fundamentale, care se referă la variabila X_n; probabilităţile fundamentale rămase, în număr de s^{n-1}, sînt atunci determinate în baza relaţiilor (XX.2,2).

Lanţul este în mod unic determinat prin cunoaşterea probabilităţilor fundamentale.

Proprietatea aceasta este evidentă, deoarece în funcţie de probabilităţile fundamentale putem calcula toate elementele lanţului, fiecare variabilă X_n fiind complet determinată din punct de vedere probabilistic.

Un lanţ general, definit prin probabilităţile lui fundamentale, se bazează pe legătura dintre o variabilă X_n şi toate variabilele anterioare $X_1, X_2, \ldots, X_{n-1}$. Prin definiţie, vom spune că *lanţul este de ordin infinit.*

In loc să ne dăm direct valorile tuturor probabilităţilor fundamentale, al căror număr, pentru variabilele de un rang destul de îndepărtat, este imens, este mai comod să ne dăm o lege de formare a probabilităţilor fundamentale referitoare la variabila X_n, în funcţie de probabilităţile fundamentale ale variabilelor anterioare.

Scanned with OKEN Scanner

Se ajunge astfel la diferite tipuri de lanțuri, a căror generalitate variază după legea de formare considerată și care sînt cuprinse sub denumirea generală de *lanțuri cu legături complete.*

Să definim cele mai simple șiruri de variabile complet înlănțuite.

Șirul

$$X_1,\ X_2,\ldots,X_n,\ldots$$

este format din variabile, care iau valorile

$$a_1,\ a_2,\ldots,a_s,\ldots$$

în număr finit.

Notăm prin

$$p_1^{(n)},\quad p_2^{(n)},\ldots,p_s^{(n)}$$

probabilitățile fundamentale ale egalităților

$$X_n=a_1,\quad X_n=a_2,\ldots,X_n=a_s,$$

care sînt date în baza faptului că cunoaștem toate valorile luate de variabilele anterioare lui X_n.

Să presupunem că variabila X_n a luat valoarea a_i. Probabilitățile fundamentale ale variabilei X_{n+1}, în baza faptului că $X_n=a_i$, devine

$$p_1^{(n+1)},\quad p_2^{(n+1)},\ldots,p_s^{(n+1)}.$$

Vom presupune atunci că

$$p_k^{(n+1)}=f_{ik}\left[p_1^{(n)},\quad p_2^{(n)},\ldots,p_s^{(n)}\right],$$

unde

$$f_{ik}\left[p_1^{(n)},\ p_2^{(n)},\ldots,p_s^{(n)}\right]$$

reprezintă o funcție dată de $p_1^{(n)},\ p_2^{(n)},\ldots,p_s^{(n)}$, îndeplinind condiția ca, pentru toate valorile $p_1^{(n)},\ p_2^{(n)},\ldots,p_s^{(n)}$ verificînd relațiile

$$0\leqslant p_1^{(n)}\leqslant 1,\quad 0\leqslant p_2^{(n)}\leqslant 1,\ldots,0\leqslant p_s^{(n)}\leqslant 1,\quad \sum_{j=1}^{s}p_j^{(n)}=1,$$

să avem

$$0\leqslant f_{ik}\left(p_1^{(n)},\quad p_2^{(n)},\ldots,p_s^{(n)}\right)\leqslant 1,$$

$$\sum_{k=1}^{s}f_{ik}\left(p_1^{(n)},\quad p_2^{(n)},\ldots,p_s^{(n)}\right)=1.$$

In baza acestor relații, rezultă că probabilitățile fundamentale ale variabilei X_{n+1} sînt funcții numai de probabilitățile fundamentale ale variabilei precedente X_n.

Lanțurile cu legături complete, definite mai sus, sînt *de ordinul întîi.* Definiția lanțurilor de ordin superior se face în mod analog.

Să presupunem astfel că probabilitățile fundamentale ale variabilelor X_n și X_{n+1} sînt respectiv

$$p_1^{(n)},\quad p_2^{(n)},\ldots,p_s^{(n)} \qquad \left(\sum_{j=1}^{s} p_j^{(n)}=1\right),$$

$$p_1^{(n+1)},\quad p_2^{(n+1)},\ldots,p_s^{(n+1)} \qquad \left(\sum_{j=1}^{s} p_j^{(n+1)}=1\right).$$

Mai admitem că variabila X_n a luat valoarea a_i, iar X_{n+1} valoarea a_j.

Vom presupune atunci că probabilitățile fundamentale $p_k^{(n+2)}$ corespunzătoare variabilei X_{n+2} sînt date prin relațiile

$$p_k^{(n+2)}=f_{ijk}(p_1^{(n)},\ p_2^{(n)},\ldots,p_s^{(n)};\ \ p_1^{(n+1)},\ p_2^{(n+1)},\ldots,p_s^{(n+1)}),$$

unde

$$0\leqq f_{ijk}(p_1^{(n)},\ p_2^{(n)},\ldots,p_s^{(n)};\ \ p_1^{(n+1)},\ p_2^{(n+1)},\ldots,p_s^{(n+1)})\leqq 1,$$

$$\sum_{k=1}^{s} f_{ijk}(p_1^{(n)},\ p_2^{(n)},\ldots,p_s^{(n)};\ \ p_1^{(n+1)},\ p_2^{(n+1)},\ldots,p_s^{(n+1)})=1,$$

pentru toate valorile

$$p_1^{(n)},\ p_2^{(n)},\ldots,p_s^{(n)},\ p_1^{(n+1)},\ p_2^{(n+1)},\ldots,p_s^{(n+1)}$$

care îndeplinesc relațiile

$$0\leqq p_1^{(n)}\leqq 1,\ 0\leqq p_2^{(n)}\leqq 1,\ldots,0\leqq p_s^{(n)}\leqq 1,\ \sum_{i=1}^{s} p_s^{(n)}=1,$$

$$0\leqq p_1^{(n+1)}\leqq 1,\ 0\leqq p_2^{(n+1)}\leqq 1,\ldots,0\leqq p_s^{(n+1)}\leqq 1,\ \sum_{i=1}^{s} p_j^{(n+1)}=1.$$

Conform definiției de mai sus, *probabilitățile fundamentale ale variabilei X_{n+2} nu depind decît de probabilitățile fundamentale ale variabilelor X_n și X_{n+1}.* Lanțul cu legături complete astfel definit este *de ordinul al doilea.*

Lanţurile de ordin mai mare se definesc pe aceeaşi cale.

Lanţurile Markov constituie un caz particular de *lanţuri cu legături complete.*

Dacă în exemplele mai sus considerate presupunem că

$$f_{ik}(p_1^{(n)}, p_2^{(n)},..., p_s^{(n)}) = p_{ik},$$

$$f_{ijk}(p_1^{(n)}, p_2^{(n)},..., p_s^{(n)}; \quad p_1^{(n+1)}, p_2^{(n+1)},..., p_s^{(n+1)}) = p_{ijk},$$

unde p_{ik}, p_{ijk} sînt constante în raport cu

$$p_1^{(n)}, p_2^{(n)},..., p_s^{(n)}, p_1^{(n+1)}, p_2^{(n+1)},..., p_s^{(n+1)},$$

obţinem categoria lanţurilor Markov, numite astfel după numele matematicianului care le-a considerat pentru prima dată.

Prin urmare, în cazul unui lanţ Markov, *probabilităţile fundamentale ale unei variabile nu depind de probabilităţile fundamentale ale variabilelor precedente, ci numai de valorile luate de aceste variabile.*

Studiul lanţurilor Markov formează un capitol însemnat în teoria probabilităţilor. Se explică acest lucru prin faptul că ele au constituit cel mai simplu exemplu de variabile dependente, iar pe de altă parte, şi-au găsit numeroase aplicaţii practice în biometrie, fizică etc.

§ 3. Diferite categorii de lanţuri

In definiţiile date pînă aici, am presupus că *funcţiile care exprimă probabilităţile fundamentale ale unei variabile în raport cu probabilităţile fundamentale ale variabilelor precedente rămîn aceleaşi oricare ar fi rangul variabilei considerate.* Ne-am referit, prin urmare, la o categorie particulară de lanţuri care, după unii autori, poartă numele de *lanţuri cu elemente constante* sau, mai simplu, *lanţuri constante,*

In cazul general, cînd *relaţiile dintre probabilităţile fundamentale variază în raport cu rangul variabilei considerate,* lanţurile reprezintă o lege de dependenţă cu mult mai complicată.

Ele sînt numite în acest caz *lanţuri cu elemente variabile.*

Scanned with OKEN Scanner

Pentru lanțul simplu, relația de bază dintre probabilitățile fundamentale devine

$$p_i^{(n+1)} = f_{ik}^{(n)} [p_1^{(n)}, p_2^{(n)}, \ldots, p_s^{(n)}],$$

prin $f_{ik}[p_1^{(n)}, p_2^{(n)}, \ldots, p_s^{(n)}]$ înțelegînd o *funcție de* $p_1^{(n)}, p_2^{(n)}, \ldots, p_s^{(n)}$ *și de indicii* i, k, n.

In mod analog se procedează și pentru lanțurile de ordin superior

Am presupus pînă acum că lanțurile admit o variabilă inițială X_1, adică formează o succesiune de forma

$$X_1, X_2, \ldots, X_n, \ldots$$

Această presupunere înseamnă o restricție destul de însemnată, pentru că în natură există numeroase exemple de lanțuri, cărora nu li se poate fixa o variabilă inițială.

Considerînd astfel un fenomen de masă, care se desfășoară în timp după legile unui lanț, din cunoașterea unui număr oarecare de variabile, apropiate de prezent, nu putem determina evoluția anterioară a lanțului de forma

$$\ldots, X_{-n}, X_{-n+1}, \ldots, X_{-1}, X_0, X_1, \ldots, X_n, \ldots$$

Lanțurile pot fi clasificate și după alte criterii.

Am presupus astfel că variabilele

$$X_1, X_2, \ldots, X_n, \ldots$$

iau valorile

$$a_1, a_2, \ldots, a_n, \ldots$$

în număr finit.

Lanțul este în acest caz *discontinuu.*

Dacă variabilele X_n sînt susceptibile de a lua toate valorile cuprinse într-un interval finit sau infinit, lanțul este *continuu.*

Să observăm că *nu este necesar ca numărul valorilor* a_i *să fie finit, pentru ca să rămînem în cazul discontinuu. Dacă, de pildă, valorile* a_i *formează o mulțime numărabilă, lanțul corespunzător este tot discontinuu.*

Definițiile din cazul continuu rămîn aceleași ca și pentru lanțurile discontinue.

Să considerăm, pentru exemplificare, cazul particular al lanțurilor lui Markov și să presupunem că variabilele lanțului pot lua valorile cuprinse într-un interval (a, b) finit sau infinit. Admițînd că X_n a

luat în acest caz valoarea x, probabilitatea ca X_{n+1} să fie cuprinsă între y şi $y+dy$ este dată de expresia $f(x, y)\,dy$.

Pentru lanţul Markov continuu şi de ordinul al doilea, presupunînd că variabilele X_n şi X_{n+1} au luat respectiv valorile x şi y, probabilitatea ca X_{n+2} să fie cuprinsă între z şi $z+dz$ este de forma $f(x, y, z)\,dz$.

In amîndouă exemplele, funcţiile $f(x, y)$ şi $f(x, y, z)$ trebuie să fie pozitive pentru toate valorile x, y, z cuprinse în intervalul (a, b) şi să îndeplinească relaţiile

$$\int_a^b f(x, y)\,dy = 1,$$

$$\int_a^b f(x, y, z)\,dz = 1.$$

Să trecem acum la cazul *mai general, al unui lanţ cu legături complete*. Notăm prin $p(t)\,dt$ probabilitatea fundamentală ca X_n să ia o valoare cuprinsă între t şi $t+dt$.

Dacă $X_n=x$, probabilitatea ca X_{n+1} să ia o valoare cuprinsă între y şi $y+dy$ este egală cu $F[x, y; p(x)]$, această expresie reprezentînd o funcţie de $p(x)$ şi x, y.

Definiţiile date se referă la lanţurile simple. Pentru lanţurile de ordin superior se procedează în mod analog.

§ 4. Legea redusă limită a sumelor de variabile înlănţuite în sensul lui Markov

Problema determinării legilor limită ale sumelor de variabile înlănţuite a fost pusă pentru prima oară de Markov [1]), care a arătat că, în cazul lanţului discontinuu, constant şi finit, legea limită coincide cu legea normală a lui Gauss-Laplace, cu condiţia de a avea $p_{ik}>0$, prin p_{ik} însemnînd probabilitatea egalităţii $x_n=a_k$ după ce $x_{n-1}=a_i$.

[1]) A. Markov, Extension des théoremes limites du calcul des probabilités à la somme des valeurs liées en chaine, Mém. Acad. Imp. Sc. Saint-Pétersbourg, 1908.

Problema pusă de Markov a fost reluată în ultima vreme cu scopul de a se înlătura ipoteza $p_{ik}>0$ şi de a găsi un mijloc de demonstraţie mai natural şi mai simplu.

Vom expune, în cazul continuu, o demonstraţie bazată pe noţiunea de funcţie caracteristică. In cazul mai simplu, al lanţurilor discontinue, rezultatele obţinute se păstrează, metoda rămînînd aceeaşi.

Plecăm de la lanţul

$$x_1,\ x_2,\ldots,\ x_n,\ldots$$

ale cărui variabile pot lua orice valoare cuprinsă în intervalul dat $(a,\ b)$.

Şirul x_n formează un lanţ Markov simplu şi constant.

Pentru definirea lui este necesar să ne dăm probabilitatea $\varphi(x,\ y)\,dy$ ca să obţinem pentru x_n o valoare cuprinsă între y şi $y+dy$, ştiind că $x_{n-1}=x$.

In mod evident, $\varphi(x,\ y)$ trebuie să îndeplinească următoarele condiţii, legate de noţiunea de probabilitate

$$\varphi(x,\ y)\geqq 0,$$

$$\int_a^b \varphi(x,\ y)\,dy=1.$$

Introducem acum probabilitatea $\Phi_n(x,\ z)\,dz$, ca suma

$$x_1+x_2+\ldots+x_n$$

să ia o valoare cuprinsă între z şi $z+dz$, ştiind că probabilitatea eventualităţii ca x_1 să ia o valoare cuprinsă între y şi $y+dy$ a fost $\varphi(x,\ y)\,dy$.

In aceleaşi condiţii, însemnăm prin $F_n(x;\ t;\ \theta)$ funcţia caracteristică a sumei

$$x_1+x_2+\ldots+x_n-n\theta,$$

unde θ este o constantă, a cărei valoare o lăsăm deocamdată nedeterminată.

In baza acestor notaţii, vom avea astfel

$$F_n(x;\ t;\ \theta)=\int_{na}^{nb} e^{(z-n\theta)\,it}\,\Phi_n(x;\ z)\,dz. \qquad \text{(XX.4,1)}$$

Pe de altă parte, probabilitățile $\Phi_n(x;\ z)\,dz$ verifică relațiile de recurență

$$\Phi_n(x;\ z)=\int_a^b \varphi(x;\ u)\,\Phi_{n-1}(u;\ z-u)\,du. \qquad \text{(XX.4,2)}$$

Ținînd seama de (XX.4,2), relația (XX.4,1) devine în acest caz

$$F_n(x;\ t;\ \theta)=\int_{na}^{nb} e^{(z-n\theta)\,it}\int_a^b \varphi(x;\ u)\,\Phi_{n-1}(u;\ z-u)\,dudz=$$

$$=\int_a^b \varphi(x;\ u)\,e^{(u-\theta)\,it}\int_{na}^{nb} e^{[z-u-(n-1)\,\theta]\,it}\,\Phi_{n-1}(u;\ z-u)\,dzdu.$$

Însă

$$\int_{na}^{nb} e^{[z-u-(n-1)\,\theta]\,it}\,\Phi_{n-1}(u;\ z-u)\,dz=\int_{na-u}^{nb-u} e^{[v-(n-1)\,\theta]\,it}\,\Phi_{n-1}(u;\ v)\,dv=$$

$$=\int_{na-u}^{nb-u} e^{[v-(n-1)\,\theta]\,it}\,\Phi_{n-1}(u,\ v)\,dv=F_{n-1}(u;\ t;\ \theta),$$

deoarece

$$na-u\leqslant(n-1)\,a\leqslant(n-1)\,b\leqslant nb-u.$$

Prin urmare

$$F_n(x;\ t;\ \theta)=\int_a^b \varphi(x;\ u)\,e^{(u-\theta)\,it}\,F_{n-1}(u;\ t;\ \theta)\,du. \qquad \text{(XX.4,3)}$$

Funcțiile caracteristice verifică deci relațiile de recurență (XX.4,3).

§ 5. Calculul asimptotic al funcțiilor caracteristice

Relațiile de recurență (XX.4,3) și faptul că

$$\int_a^b \varphi(x;\ u)\,du=1$$

conduc în mod natural la ideea de a considera ecuația integrală

$$F(x;\ t;\ \theta)=\lambda\int_a^b \varphi(x;\ u)\,e^{(u-\theta)\,it}\,F(u;\ t;\ \theta)\,du+1. \qquad \text{(XX.5,1)}$$

Scanned with OKEN Scanner

Să presupunem că $\varphi(x;\ u)$ este mărginită în pătratul (D)

$$a \leqq x \leqq b, \quad a \leqq u \leqq b.$$

Avem cu alte cuvinte

$$|\varphi(x;\ u)| < M,$$

unde M este o constantă dată, pentru toate valorile lui x și u cuprinse între marginile a și b inclusiv.
Rezultă atunci că

$$|\varphi(x;u)\, e^{(u-\theta)\,it}| < |\varphi(x;u)| < M.$$

In aceste condiții, dacă λ nu este o rădăcină a ecuației

$$D(\lambda;t;\theta) =$$

$$= 1 + \sum_{p=1}^{\infty} \frac{(-\lambda)^p}{p!} \int_a^b \int_a^b \cdots \int_a^b e^{(s_1+s_2+\ldots+s_p-p\theta)\,it}\, K\begin{pmatrix} s_1 & s_2 \ldots s_p \\ s_1 & s_2 \ldots s_p \end{pmatrix} ds_1\, ds_2 \ldots ds_p, \tag{XX.5,2}$$

se știe după o cunoscută teoremă a lui Fredholm că ecuația integrală (XX.5,1) admite soluția unică

$$F(x;t;\theta) = 1 + \int_a^b \frac{D\left[\begin{matrix} x \\ s \end{matrix}\middle|\, \lambda;t;\theta\right]}{D[\lambda;t;\theta]}\, ds, \tag{XX.5,3}$$

unde

$$D\left[\begin{matrix} x \\ s \end{matrix}\middle|\, \lambda;t;\theta\right] = \varphi(x;y)\, e^{(y-\theta)\,it} + \sum_{p=1}^{\infty} \frac{(-\lambda)^p}{p!} \int_a^b \int_a^b \cdots$$

$$\ldots \int_a^b e^{[y+s_1+s_2+\ldots+s_p-(p+1)\,\theta]\,it}\, K\begin{pmatrix} x & s_1 & s_2 \ldots s_p \\ y & s_1 & s_2 \ldots s_p \end{pmatrix} ds_1\, ds_2 \ldots ds_p, \tag{XX.5,4}$$

iar $K\begin{pmatrix} x & s_1 & s_2 \ldots s_p \\ y & s_1 & s_2 \ldots s_p \end{pmatrix}$ este determinantul clasic al lui Fredholm.
Pe de altă parte, soluția (XX.5,3) poate fi pusă și sub forma

$$F(x;t;\theta) = 1 + \lambda F_2(x;t;\theta) + \ldots + \lambda^n F_n(x;t;\theta) + \ldots, \tag{XX.5,5}$$

care se deduce din relațiile de recurență (XX.4,3).

Faptul că $F_n(x;t;\theta)$ este coeficientul lui λ^n în dezvoltarea soluției ecuației integrale (XX.5,1) după puterile lui λ ne va permite să determinăm valoarea asimptotică a funcției caracteristice.

Pentru aceasta, menționăm că ecuația integrală omogenă, care corespunde nucleului $\varphi(x, u)$

$$F(x;\theta)=\lambda\int_a^b \varphi(x, u)\ F(u, \theta)\,dt, \qquad \text{(XX.5,6)}$$

unde θ este o constantă, a fost considerată pentru prima oară de J. Hadamard (Congresul matematicienilor de la Bologna, 1928, tomul V, pp. 133—139), care a arătat că toate constantele fundamentale ale nucleului $\varphi(x;u)$ sînt mai mari în modul decît unitatea.

Demonstrația acestei proprietăți este analogă cu aceea care ne-a condus, în cazul discontinuu, la concluzia că rădăcinile ecuației caracteristice sînt mai mici în modul decît 1. Pentru acest motiv, nu o redăm aici.

In același mod se demonstrează și proprietatea că orice constantă fundamentală de modul 1 este rădăcina unei ecuații binome.

Dacă în ecuația integrală omogenă

$$F(x;t;\theta)=\lambda\int_a^b \varphi(x;u)\,e^{(u-\theta)it}\,F(u;t;\theta)\,du \qquad \text{(XX.5,7)}$$

facem $t=0$, obținem ecuația omogenă (XX.5,6) studiată de J. Hadamard. Și pentru această ecuație, repetînd aceeași demonstrație, se arată că toate constantele fundamentale, corespunzătoare nucleului $\varphi(x;u)\,e^{(u-\theta)it}$ sînt mai mari în modul decît 1.

Mulțimile finale și de trecere, considerate în cazul lanțurilor discontinue, capătă definiții identice și pentru lanțurile continue.

Vom considera, pentru simplificare, cazul cînd lanțul admite o singură mulțime finală; rezultă, drept consecință, întocmai ca și în cazul discontinuu, că ecuația $D(\lambda;0;\theta)=1$ admite pe $y=1$ ca rădăcină simplă.

Scanned with OKEN Scanner

Dacă funcția întreagă $D(\lambda; t; \theta)$ nu se anulează pentru $\lambda = \infty$, descompunerea funcției meromorfe

$$\frac{D\left[\begin{matrix} x \\ y \end{matrix} \middle| t; \lambda; \theta\right]}{D(\lambda; t; \theta)}$$

în elemente simple nu conține o parte întreagă. Deoarece $\lambda_j(t)$ este un pol de ordinul p, termenul relativ la acest pol în descompunerea lui $F(x; t; \theta)$ va fi de forma

$$\Gamma_j(\lambda) = \frac{A_{j,1}(t; x)}{1 - \frac{\lambda(t)}{\lambda_j(t)}} + \frac{A_{j,2}(t; x)}{\left[1 - \frac{\lambda(t)}{\lambda_j(t)}\right]^2} + \ldots + \frac{A_{j,p}(t; x)}{\left[1 - \frac{\lambda(t)}{\lambda_j(t)}\right]^p} =$$

$$= \sum_{n=0}^{\infty} \frac{\lambda^n(t)}{\lambda_j^n(t)} A_{j,1}(t; x) + \frac{n+1}{1!} A_{j,2}(t; x) + \ldots$$

$$\ldots + \frac{(n+1)(n+2)\ldots(n+p+1)}{(p-1)!} A_{j,p}(t; x) = \sum_{n=0}^{\infty} \frac{1}{n} P_j(n; x; t) \lambda^n,$$

$P_j(n; x; t)$ fiind un polinom în n de gradul maxim $(p-1)$. Dacă vom considera toți polii funcției de mai sus, prin identificarea coeficientului lui λ^n rezultat din operația precedentă cu (XX.5,5), obținem

$$F_n(x; t; \theta) = \sum_j \frac{1}{\lambda_j^n(t)} R_j(n; t; x; \theta), \qquad \text{(XX.5,8)}$$

unde $R_j(n; t; x; \theta)$ este un polinom în n, depinzînd în mod evident și de constanta θ, a cărui existență în fiecare dintre relațiile de mai sus nu este înlăturată, chiar dacă acest lucru nu reiese din notație.

Egalitatea (XX.5,8) a fost obținută în ipoteza că $D(\lambda; t; \theta)$ nu se anulează o dată cu $\frac{s}{\lambda}$. Această relație este adevărată însă și cînd această ipoteză nu se verifică, cu condiția ca valorile lui n să fie mai mari decît exponentul celei mai mari puteri a lui λ în partea întreagă care figurează în dezvoltarea lui $\frac{D\left[\begin{matrix} x \\ y \end{matrix} \middle| \lambda; t; \theta\right]}{D(\lambda; t; \theta)}$.

Scanned with OKEN Scanner

$F_n(x;t;\theta)$ fiind o funcție caracteristică, trebuie să avem $F_n(x;t;\theta)=1$; din (XX.5,8) rezultă astfel

$$\sum_j \frac{1}{\lambda_j^n(0)} R_j[n;x;0;\theta]=1, \qquad \text{(XX.5,9)}$$

oricare ar fi n.

Să facem acum o discriminare a diferitelor cazuri ce se pot întîmpla.

Rădăcinile $\lambda_j(0)$ am arătat că sînt toate mai mari sau egale în modul cu unu, iar cele egale în modul cu unitatea verifică ecuația binomă

$$\lambda^s=1,$$

unde s este un număr întreg și pozitiv.

Fie atunci

$$\lambda_1(t;\theta),\ \lambda_2(t;\theta),\ \ldots,\ \lambda_s(t;\theta) \qquad \text{(XX.5,10)}$$

constantele fundamentale ale nucleului $\varphi(x;u)\,e^{(u-\theta)\,it}$ care, pentru $t\to 0$, tind către rădăcinile ecuației binome de mai sus.

Celelalte constante fundamentale fiind mai mari în modul decît 1, constituie în primul membru al relației (XX.5,9) un termen $H(n)$ care tinde către zero, cînd n crește necontenit.

Rezultă atunci din (XX.5,9) că

$$\sum_{j=1}^{s} \frac{1}{\lambda_j^n(0)} R_j(x;0;\theta)=1-H(n), \qquad \text{(XX.5,11)}$$

suma fiind extinsă la rădăcinile (XX.5,10) pentru $t=0$, iar

$$\lim_{n\to\infty} H(n)=0.$$

In formula (XX.5,11) trebuie să observăm că, în loc de polinomul $R_j(n;x;0;\theta)$ în raport cu n, am trecut constanta $R_j(x;0;\theta)$, deoarece rădăcinile $\lambda(0)$ sînt simple.

Cum $\lambda_1(0)=1$, relația (XX.5,11) se mai poate scrie

$$[R_1(x;0;\theta)-1]\,\lambda_1^{-n}(0)+R_2(x;0;\theta)\,\lambda_2^{-n}(0)+\ldots$$
$$\ldots+R_s(x;0;\theta)\,\lambda_s^{-n}(0)=-H_n,$$

egalitate care trebuie să rămînă adevărată pentru toate valorile întregi date lui n.

Membrul al doilea al acestei relații tinde către zero, în timp ce membrul întîi oscilează în jurul a s valori finite; ajungem astfel la un sistem de s ecuații

$$[R_1(x;0;\theta)-1]\lambda_1^{-n}(0)+R_2(x;0;\theta)\lambda_2^{-n}(0)+\ldots$$

$$\ldots+R_s(x;0;\theta)\lambda_s^{-n}(0)=0,$$

al căror determinant este diferit de zero. Prin urmare, trebuie să avem

$$R_1(x;0;\theta)=1,\ R_2(x;0;\theta)=0,\ldots,R_s(x;0;\theta)=0. \quad \text{(XX.5,12)}$$

§ 6. Momentul de ordinul întîi

Ne putem alege pe t destul de mic, astfel încît rădăcinile (XX.5,10) să fie simple, întocmai ca și limitele lor pentru $t=0$. Putem atunci scrie relația (XX.5,8) sub forma

$$F_n(x;t;\theta)=\sum_{j=1}^{s}R_j(x;t;\theta)\,\mu_j^n(t;\theta)+\sum_{g}R_g(n;x;t;\theta)\,\mu_j^n(t;\theta), \quad \text{(XX.6,1)}$$

prima sumă fiind extinsă la rădăcinile (XX.5,10), a doua la restul rădăcinilor, iar

$$\mu_h(t;\theta)=\frac{1}{\lambda_h(t;\theta)}.$$

Prin derivare obținem

$$\frac{\partial F_n(x;t;\theta)}{\partial t}=\sum_{j=1}^{s}\frac{\partial R_j(x;t;\theta)}{\partial t}\mu_j^n(t;\theta)+$$

$$+n\sum_{j=1}^{s}R_j(x;t;\theta)\,\mu_j^{n-1}(t;\theta)\,\mu_j'(t;\theta)+G_n(t;\theta), \quad \text{(XX.6,2)}$$

$G_n(t;\theta)$ fiind o funcție care tinde către zero o dată cu t.
Din definiția funcției caracteristice se știe însă că

$$\left[\frac{\partial F_n(x;t;\theta)}{\partial t}\right]_{t=0}=M^{(1)}(x;n;\theta),$$

prin $M^{(1)}(x;n;\theta)$ înţelegînd momentul de ordinul întîi al sumei

$$x_1+x_2+\ldots+x_n-n\theta.$$

Făcînd atunci pe $t=0$, din (XX.6,2) obţinem

$$M^{(1)}(x;n;\theta)=n\mu_1'(0;\theta)+E(x;\theta;n), \qquad \text{(XX.6,3)}$$

$E(x;\theta;n)$ fiind o funcţie asimptotică periodică de n. Să observăm faptul important că termenul $\mu_1'(0,\theta)$, care reprezintă coeficientul lui n în $M^{(1)}(x;n;\theta)$ nu depinde de x.
Putem astfel scrie

$$\lim_{n\to\infty}\frac{M^{(1)}(x;n;\theta)}{n}=\mu_1'(0;\theta),$$

adică tocmai teorema lui Fréchet din cazul discontinuu.
Să admitem că există o valoare $\theta=a$, pentru care

$$\lambda_1'(0;a)=0.$$

Un calcul simplu, pe care este inutil să-l mai reproducem, ne arată că în acest caz avem şi

$$\lambda_j'(0;a)=0$$

pentru toţi indicii corespunzători valorilor (XX.5,8).
Pe de altă parte, valoarea lui a se obţine imediat derivînd egalitatea (XX.6,1) în raport cu t şi făcînd apoi $t=0$.
Rezultă, astfel, ţinînd seama că $\lambda_1(0;a)=1$, $\lambda_1'(0;a)=0$

$$a=\frac{\sum_{p=1}^{\infty}\frac{(-1)^p}{p!}\int_a^b\int_a^b\cdots\int_a^b(s_1+s_2+\ldots+s_p)\,K\begin{pmatrix}s_1,\ s_2,\ \ldots,\ s_p\\ s_1,\ s_2,\ \ldots,\ s_p\end{pmatrix}ds_1\,ds_2\ldots ds_p}{\sum_{p=1}^{\infty}\frac{(p-1)!}{(-1)^p}\int_a^b\int_a^b\cdots\int_a^b K\begin{pmatrix}s_1,\ s_2,\ \ldots,\ s_p\\ s_1,\ s_2,\ \ldots,\ s_p\end{pmatrix}ds_1\,ds_2\ldots ds_p}. \qquad \text{(XX.6,4)}$$

§ 7. Valoarea asimptotică a momentului de ordinul al doilea

Intocmai ca în cazul momentului de ordinul întîi, vom deriva relația (XX.6,1) de două ori în raport cu t şi vom face apoi $t=0$. Vom obține astfel

$$\left[\frac{\partial^2 F_n(x;t;a)}{\partial t^2}\right]_{t=0} = -\lambda_1''(0;a)\,n+$$

$$+\sum_{j=2}^{s}\left[\frac{\partial^2 R_j(x;t;a)}{\partial t^2}\right]_{t=0}\mu_j^n(0;a)+L_n(x;a),$$

unde

$$\lim_{n\to\infty} L_n(x;a)=0.$$

Dacă însemnăm, prin urmare, prin $M^{(2)}(x;n;a)$ momentul de ordinul al doilea al sumei $x_1+x_2+\ldots+x_n-n\theta$, avem

$$M^{(2)}(x;n;\theta)\simeq -\lambda_1''(0;a)\,n+\sum_{j=2}^{s}\left[\frac{\partial^2 R_j(x;t;a)}{\partial t^2}\right]_{t=0}\mu_j^n(0;a).$$

Este de observat că şi în această formulă, x nu intervine în coeficientul lui n din membrul al doilea.

Cît pentru valoarea lui $\lambda_1''(0;a)$, ea se determină imediat din egalitatea

$$\sum_{p=0}^{\infty}\frac{(-\lambda_1)^p}{p!}\int_a^b\int_a^b\ldots\int_a^b e^{(s_1+s_2+\ldots+s_p-pa)\,it}K\begin{pmatrix}s_1, & s_2, \ldots, & s_p\\ s_1, & s_2, \ldots, & s_p\end{pmatrix}ds_1\,ds_2\ldots ds_p=0$$

pe care o derivăm de două ori în raport cu t şi ţinem seama de faptul că $\lambda_1(0)=1$, $\lambda_1'(0)=0$.

Obţinem astfel

$$-\lambda_1''(0;\,a)=$$

$$=\frac{\sum_{p=0}^{\infty}\frac{(-1)^p}{p!}\int_a^b\int_a^b\ldots\int_a^b e^{(s_1+s_2+\ldots+s_p-a)}K\begin{pmatrix}s_1, & s_2, \ldots, & s_p\\ s_1, & s_2, \ldots, & s_p\end{pmatrix}ds_1\,ds_2\ldots ds_p}{\sum_{p=0}^{\infty}\frac{(-1)^p}{(p-1)!}\int_a^b\int_a^b\ldots\int_a^b K\begin{pmatrix}s_1, & s_2, \ldots, & s_p\\ s_1, & s_2, \ldots, & s_p\end{pmatrix}ds_1\,ds_2\ldots ds_p}.$$

Scanned with OKEN Scanner

§ 8. Comportarea asimptotică a funcțiilor caracteristice

Momentul de ordinul al doilea $M^{(2)}(x; n; \theta)$ fiind, din punct de vedere asimptotic, de ordinul lui n, pentru a obține la limită o lege redusă se impune transformarea $t = \frac{u}{\sqrt{n}}$.

Ținînd seama de (XX.5,12), rezultă atunci din (XX.6,1)

$$\lim_{n\to\infty} F_n\left(x; \frac{u}{\sqrt{n}}; a\right) = \lim_{n\to\infty} \mu_1^n\left(\frac{u}{\sqrt{n}}; a\right).$$

Insă

$$\mu_1\left(\frac{u}{\sqrt{n}}; a\right) = 1 + \frac{\mu_1''(0; a)}{2} \cdot \frac{u^2}{n} + \ldots = 1 - \frac{\lambda_1''(0; a}{2} \cdot \frac{u^2}{n} + \ldots - \mu^2.$$

Rezultă astfel

$$\lim_{n\to\infty} F_n\left(x; \frac{u}{\sqrt{n}}; a\right) = e^{-\frac{\lambda_1''(0; a)}{2}\mu^2}$$

sau

$$\lim_{n\to\infty} F_n\left(x; \frac{u}{\sqrt{n\lambda_1''(0; a)}}; a\right) = e^{-\frac{u^2}{2}}.$$

Membrul al doilea reprezintă, însă, tocmai funcția caracteristică a legii normale a lui Laplace.

Reiese astfel că, *dacă* $\lambda_1''(0; a) \neq 0$, *legea redusă a variabilei*

$$x_1 + x_2 + \ldots + x_n,$$

pentru n *crescînd necontenit, tinde către legea normală a lui Laplace.*

§ 9. Cazul lanțului Markov finit

Considerațiile de mai sus, referitoare la lanțul Markov continuu, simplu și constant, sînt valabile și pentru lanțul Markov finit, simplu și constant.

Fie

$$X_n = \begin{pmatrix} a_1, & a_2, \ldots, & a_n \\ p_{i1}, & p_{i2}, \ldots, & p_{in} \end{pmatrix} \quad (n=1, 2, \ldots) \qquad \text{(XX.9,1)}$$

un şir infinit de variabile aleatorii, unde p_{ik} este probabilitatea de trecere de la valoarea a_i la valoarea a_{ik}.

Să notăm cu $F(n; t; \theta)$ funcţia caracteristică a sumei $X_{(n)} = X_1 + X_2 + \ldots + X_n - n\theta$, unde θ este o constantă oarecare; să notăm de asemenea cu $F_j(n; t; \theta)$ funcţia caracteristică a lui $X_{(n)}$ în ipoteza că

$$X_1 = \begin{pmatrix} a_1, & a_2, \ldots, & a_m \\ p_{j1}, & p_{j2}, \ldots, & p_{im} \end{pmatrix}$$

şi cu $F_{jk}(n; t; \theta)$ funcţia caracteristică a lui $X_{(n)}$ în ipoteza că

$$X_1 = \begin{pmatrix} a_1, & a_2, \ldots, & a_m \\ p_{j1}, & p_{j2}, \ldots, & p_{jm} \end{pmatrix}$$

şi

$$X_n = a_k.$$

In ceea ce urmează, t este variabila specifică funcţiei caracteristice. Fie $P_k(n; \lambda_1, \lambda_2, \ldots, \lambda_m)$ probabilitatea ca $X_{(n)}$ să ia valoarea

$$\lambda_1 a_1 + \lambda_2 a_2 + \ldots + \lambda_m a_m - n\theta,$$

unde $\lambda_1, \lambda_2, \ldots, \lambda_m$ sînt întregi pozitivi sau nuli astfel încît

$$\lambda_1 + \lambda_2 + \ldots + \lambda_m = n.$$

Chiar prin definiţie, valoarea funcţiei caracteristice va fi

$$F_k(n, t, \theta) = \sum_{\lambda_1, \lambda_2, \ldots, \lambda_m} P_k(n; \lambda_1, \lambda_2, \ldots, \lambda_m) \times$$

$$\times e^{(\lambda_1 a_1 + \lambda_2 a_2 + \ldots + \lambda_m a_m - n\theta)\, it}, \qquad \text{(XX.9,2)}$$

unde sumarea se referă la toate valorile $\lambda_1, \lambda_2, \ldots, \lambda_m$ întregi pozitive sau nule, $\lambda_1 + \lambda_2 + \ldots + \lambda_m = n$.

Scanned with OKEN Scanner

Pe de altă parte,

$$P_k(n;\ \lambda_1,\ \lambda_2,\ldots,\ \lambda_m)=$$

$$=\sum_{j=1}^{m} p_{kj} P_j(\lambda_1,\ \lambda_2,\ldots,\ \lambda_{j-1},\ \lambda_j-1,\ \lambda_{j+1},\ldots,\ \lambda_m). \quad \text{(XX.9,3)}$$

Dacă utilizăm această relație în egalitatea (XX.9,2), obținem

$$F_k(n;\ t;\ \theta)=\sum_{j=1}^{m} p_{kj} P_j(\lambda_1,\ \lambda_2,\ldots,\ \lambda_{j-1},\ \lambda_j-1,\ \lambda_{j+1},\ldots,\ \lambda_m)\times$$

$$\times e^{(\lambda_1 a_1+\lambda_2 a_2+\ldots+\lambda_m a_m-n\theta)\,it}$$

sau încă

$$F_k(n;\ t;\ \theta)=$$

$$=\sum_{\lambda_1,\lambda_2,\ldots,\lambda_m} p_{kj}\, e^{(a_j-\theta)\,it} P_j(\lambda_1,\ \lambda_2,\ldots,\lambda_{j-1},\ \lambda_j-1,\ \lambda_{j+1},\ldots,\ \lambda_m)\times$$

$$\times e^{[\lambda_1 a_1+\ldots+\lambda_{j-1} a_{j-1}+\ldots+\lambda_m a_m-(n-1)\theta]\,it}.$$

Deoarece, pe de altă parte,

$$\sum_{\lambda_1,\lambda_2,\ldots,\lambda_m} P_j(\lambda_1,\ \lambda_2,\ldots,\ \lambda_{j-1},\ \lambda_j-1,\ \lambda_{j+1},\ldots,\ \lambda_m)\times$$

$$\times e^{[\lambda_1 a_1+\ldots+\lambda_{j-1} a_{j-1}+(\lambda_j-1) a_j+\lambda_{j+1} a_{j+1}+\ldots+\lambda_m a_m-(n-1)\theta]\,it}=$$

$$=F_j(n-1;\ t;\ \theta),$$

din egalitatea precedentă rezultă relația fundamentală

$$F_k(n;\ t;\ \theta)=\sum_{j=1}^{m} p_{kj} e^{(a_j-\theta)\,it} F_j(n-1;\ t;\ \theta), \qquad \text{(I)}$$

pe care o putem scrie și sub forma

$$F_k(n;\ t;\ \theta)=\sum_{j=1}^{m} F_{kj}(1;\ t;\ \theta)\, F_j(n-1;\ t;\ \theta).$$

Relația (I) ne arată că funcțiile caracteristice $F_k(n;\ t;\ \theta)$ $(k=1,\ 2,\ldots,\ m)$ sînt iteratele de ordinul n ale valorilor $F_k(1;\ t;\ \theta)$ prin transformarea liniară

$$x'_k=\sum_{j=1}^{m} a_{kj}\, x_j,$$

$$a_{kj}=p_{kj}\, e^{(a_j-\theta)\,it}.$$

Să notăm cu A matricea $\| a_{kj} \|$, cu E matricea unitate cu m linii și m coloane, cu $\lambda_1, \lambda_2, \ldots, \lambda_s$ cele s rădăcini distincte ale ecuației caracteristice a matricei

$$\| A - \lambda E \|$$

și cu $m_1, m_2, \ldots, m_s$ ordinele de multiplicitate respective. Rezultă atunci

$$|A - \lambda E| = (\lambda - \lambda_1)^{m_1} (\lambda - \lambda_2)^{m_2} \ldots (\lambda - \lambda_s)^{m_s}.$$

Dacă punem

$$a_{rs}^{(n)} = \sum_{p=1}^{m} a_{rp} a_{ps}^{(n-1)} = \sum_{p=1}^{m} a_{rp}^{(n-1)} a_{ps},$$

cu

$$a_{rs}^{(1)} = a_{rs},$$

și dacă notăm încă cu $G_{rs}(\lambda)$ minorul elementului a_{rs} în determinantul $|A - \lambda E|$, vom avea, după o demonstrație a lui O. Perron,

$$a_{rs}^{(n)} = \sum_{j=1}^{s} \frac{1}{(m_j - 1)!} D_{\lambda}^{m_j - 1} \left[\frac{\lambda^n G_{rs}(\lambda)}{\Psi_j(\lambda)} \right]_{\lambda = \lambda_j}, \qquad \text{(XX.9,4)}$$

unde D_{λ}^{p} reprezintă operatorul de derivare de ordinul p în raport cu λ și

$$\psi_j(\lambda) = \frac{|A - \lambda E|}{(\lambda - \lambda_j)^{m_j}}.$$

Expresia (XX.9,4) dă soluția problemei care ne interesează. Într-adevăr, relația (I) poate fi scrisă

$$F_k(n;\ t;\ \theta) = \sum_{j=1}^{m} a_{kj} F_j(n-1;\ t;\ \theta)$$

și, în virtutea aceleiași relații,

$$F_j(n-1;\ t;\ \theta) = \sum_{j=1}^{m} a_{jh} F_h(n-2;\ t;\ \theta);$$

obținem

$$F_k(n;\ t;\ \theta) = \sum_{h=1}^{m} a_{kh}^{(n-1)} F_h(1;\ t;\ \theta).$$

Pe de altă parte,

$$F_k(1;\ t;\ \theta)=\sum_{i=1}^{m} p_{hj}e^{(a_j-\theta)\,it}=\sum_{j=1}^{m} a_{hj};$$

deci

$$F_k(n;\ t;\ \theta)=\sum_{j=1}^{m} a_{kh}^{(n-1)}a_{hj}=\sum_{j=1}^{m} a_{kj}^{(n)}.$$

Folosind acum formula (XX.9,4), obținem expresia căutată

$$F_k(n;\ t;\ \theta)=\sum_{h=1}^{m}\sum_{j=1}^{s}\frac{1}{(m_j-1)!}D_{\lambda}^{m_j-1}\left[\frac{\lambda^n G_{kh}(\lambda;\ t)}{\Psi_j(\lambda)}\right]_{\lambda=\lambda_j}. \qquad \text{(II)}$$

Această expresie ne va servi atît pentru calculul direct al formei asimptotice a lui $F_k(n;\ t;\ \theta)$ cînd $n\to\infty$, cît și pentru calculul momentelor de diferite ordine ale variabilei $X(n)$.

§ 10. Rădăcinile ecuației caracteristice

Să punem $D(\lambda;\ t;\ \theta)=|A-\lambda E|$. Atunci

$$D(\lambda;\ t;\ \theta)=\begin{Vmatrix} \lambda-p_{11}e^{(a_1-\theta)\,it} & -p_{12}e^{(a_2-\theta)\,it} & \dots & -p_{1m}e^{(a_m-\theta)\,it} \\ -p_{21}e^{(a_1-\theta)\,it} & \lambda-p_{22}e^{(a_2-\theta)\,it} & \dots & -p_{2m}e^{(a_m-\theta)\,it} \\ \dots & \dots & \dots & \dots \\ -p_{m1}e^{(a_1-\theta)\,it} & -p_{m2}e^{(a_2-\theta)\,it} & \dots & \lambda-p_{mm}e^{(a_m-\theta)\,it} \end{Vmatrix}. \qquad \text{(III)}$$

Rădăcinile $\lambda_l=\lambda_l(t;\ \theta)$ ale ecuației $D(\lambda;\ t;\ \theta)=0$ și gradele lor de multiplicitate respective $m_l=m_l(t;\ \theta)$ sînt, în general, funcții de t și θ.

Vom demonstra mai întîi că avem întotdeauna

$$|\lambda_l(t;\ \theta)|\leqslant 1$$

pentru orice t. Procedăm în acest scop ca în cazul cînd ar fi vorba de ecuația

$$D(\lambda;\ 0;\ \theta)=0$$

bine cunoscută în teoria lanțurilor Markov.

Pentru orice rădăcină $\lambda_g(t;\ \theta)$ putem găsi întotdeauna m valori $Z_1,\ Z_2,\ldots,\ Z_m$ pentru care

$$\sum_{h=1}^{m} p_{jh} e^{(a_h-\theta)\,it} Z_h = \lambda_g(t;\ \theta)\, Z_j. \qquad \text{(XX.10,1)}$$

De aici rezultă

$$\sum_{h=1}^{m} p_{jh} |Z_h| \geqq |\lambda_g(t;\ \theta)|\,|Z_j|. \qquad \text{(XX.10,2)}$$

Fie $|Z_j|$ cel mai mare dintre modulele $|Z_h|$ $(h=1,\ 2,\ldots,\ m)$. Inegalitatea precedentă ne dă atunci inegalitatea

$$|\lambda_g(t;\ \theta)| \leqq \sum_{h=1}^{m} p_{jh} \frac{|Z_h|}{|Z_j|} \leqq \sum_{h=1}^{m} p_{jh} = 1,$$

ceea ce demonstrează afirmația noastră.

Putem merge mai departe în această cercetare, examinînd cazurile în care

$$|\lambda_g(t;\ \theta)| = 1. \qquad \text{(XX.10,3)}$$

In aceste cazuri, în locul relației (XX.10,2) avem

$$\sum_{h=1}^{m} p_{jh} |Z_h| = |Z_j|,$$

de unde rezultă

$$|Z_1| = |Z_2| = \ldots = |Z_m|.$$

Să punem atunci

$$\lambda_g(t;\ \theta) = e^{i\varphi_g(t;\ \theta)},$$

$$Z_h = \rho e^{i\omega_h(t;\ \theta)} \quad (h=1,\ 2,\ldots,\ m).$$

Cu aceste notații, din ecuația (XX.10,1) obținem

$$\sum_{h=1}^{m} p_{jh} e^{(a_h-\theta)\,it} e^{i\omega_h(t;\ \theta)} = e^{i\varphi_g(t;\ \theta)} e^{i\omega_j(t;\ \theta)} \quad (j=1,\ 2,\ldots,\ m),$$

de unde rezultă ecuațiile în termeni reali

$$\sum_{h=1}^{m} p_{jh} \cos[(a_h-\theta)\,t + \omega_h(t;\ \theta)] = \cos[\omega_j(t;\ \theta) + \varphi_g(t;\ \theta)],$$

$$\sum_{h=1}^{m} p_{jh} \sin[(a_h-\theta)t+\omega_h(t;\ \theta)]=\sin[\omega_j(t;\ \theta)+\varphi_g(t;\ \theta)].$$

Rezultă, de aici, imediat ecuația

$$\sum_{h=1}^{m} p_{jh} \cos[(a_h-\theta)t+\omega_h-\omega_j-\varphi_g]=1. \qquad \text{(XX.10,4)}$$

Deoarece $p_{jh}\geqq 0$ și $\sum_{h=1}^{m} p_{jh}=1$, din relația (XX.10,4) reiese că, pentru toate valorile lui h pentru care $p_{jh}\neq 0$, trebuie să avem

$$(a_h-\theta)t+\omega_h-\omega_j-\varphi_g=2k_{jh}\pi,$$

k_{ij} fiind un întreg pozitiv sau negativ.

Să considerăm unul dintre șirurile de valori $a_{h_1}, a_{h_2}, \ldots, a_{h_s}, a_{h_1}$ care pot fi luate succesiv de variabilele lanțului, astfel ca toate probabilitățile de trecere corespunzătoare să fie diferite de zero

$$p_{h_1h_2}\neq 0,$$
$$p_{h_2h_3}\neq 0, \ldots, p_{h_{s-1}h_s}\neq 0,$$
$$p_{h_sh_1}\neq 0.$$

Din egalitatea (II) obținem atunci

$$\begin{aligned}
&(a_{h_2}-\theta)t+\omega_{h_2}-\omega_{h_1}-\varphi_g=2k_{12}\pi,\\
&(a_{h_3}-\theta)t+\omega_{h_3}-\omega_{h_2}-\varphi_g=2k_{23}\pi,\\
&\cdots\cdots\cdots\cdots\cdots\cdots\\
&(a_{h_s}-\theta)t+\omega_{h_s}-\omega_{h_{s-1}}-\varphi_g=2k_{s-1,\,s}\pi,\\
&(a_{h_1}-\theta)t+\omega_{h_1}-\omega_{h_s}-\varphi_g=2k_{s1}\pi,
\end{aligned}$$

de unde, prin adunare,

$$\left(\sum_{j=1}^{s} a_{h_j}-s\theta\right)t-s\varphi_g=2k\pi,$$

ceea ce ne arată că

$$\varphi_g=\left(\frac{a_{h_1}+a_{h_2}+\ldots+a_{h_s}}{s}-\theta\right)t+\frac{2k\pi}{s}. \qquad \text{(XX.10,5)}$$

Scanned with OKEN Scanner

Pentru $t=0$ avem $\varphi_g=\frac{2k\pi}{s}$, k și s fiind întregi. Deci rădăcina $\lambda_g=e^{\varphi g i}$, al cărei modul este egal cu unitatea, este o rădăcină a ecuației binome $x^s-1=0$, rezultat bine cunoscut în teoria lanțurilor Markov.

Egalitatea (XX.10,5) este valabilă pentru fiecare dintre șirurile de valori de tipul precedent; dacă $a_{j_1}, a_{j_2}, \ldots, a_{j_\sigma}$ este un al doilea șir, atunci

$$\left(\frac{a_{h_1}+a_{h_2}+\ldots+a_{h_s}}{s}-\theta\right)t+\frac{2k\pi}{s}=\left(\frac{a_{j_1}+a_{j_2}+\ldots+a_{j_\sigma}}{s}-\theta\right)t+\frac{2k\pi}{\sigma}$$

pentru orice t. Deci

$$\frac{a_{h_1}+a_{h_2}+\ldots+a_{h_s}}{s}=\frac{a_{j_1}+a_{j_2}+\ldots+a_{j_\sigma}}{\sigma}. \qquad \text{(XX.10,6)}$$

Condițiile exprimate de (XX.10,6) pentru orice sistem de forma $a_{h_1}, a_{h_2}, \ldots, a_{h_s}, a_{h_1}$ sînt de asemenea sufiçiente pentru ca

$$\lambda_g(t;\theta)=1 \qquad \text{(XX.10,7)}$$

pentru orice t.

Fie

$$D(\lambda;t;\theta)=\lambda^m+B_1^{(t)}\lambda^{m-1}+\ldots+B_l^{(t)}\lambda^{m-l}+\ldots+B_m^{(t)}=0$$

dezvoltarea determinantului $D(\lambda;t;\theta)$.

Să observăm că

$$1+B_1(0)+B_2(0)+\ldots+B_m(0)=0, \qquad \text{(XX.10,8)}$$

deoarece $\lambda=1$ este o rădăcină a ecuației $D(\lambda;t;\theta)=0$.

Dacă punem

$$p_{kj}e^{(a_j-\theta)it}=a_{kj},$$

coeficientul B_{m-s} al lui λ^s va avea forma

$$B_{m-s}=\sum(\pm a_{\alpha\alpha'}a_{\beta\beta'}\ldots a_{\mu\mu'}),$$

unde $\alpha, \beta, \ldots, \mu, \alpha', \beta', \ldots, \mu'$ sînt întregi pozitivi cuprinși între 1 și m, într-o ordine oarecare, numărul factorilor fiind $m-s$.

Pentru a ne da seama de structura termenilor care compun coeficientul B_{m-s}, să considerăm termenul corespunzător lui λ^s, format

din primele linii și coloane. Intregii $\alpha, \beta, \ldots, \mu$ trebuie căutați printre valorile $s+1, s+2, \ldots, m$.

Să plecăm de la valoarea $a_{\alpha\alpha'}$; dat fiind că cele $m-s-1$ valori $\beta, \gamma, \ldots, \mu$ trebuie căutate printre cele $m-s-1$ numere $s+1, s+2, \ldots, m$ (afară de α), vom găsi un număr egal cu α', afară de cazul $\alpha'=\alpha$. Vom da astfel lui $a_{\alpha\alpha'}$ valoarea $a_{\alpha'\alpha''}$ etc. Ajungem astfel la un produs de forma

$$a_{\alpha\alpha'}a_{\alpha'\alpha''}a_{\alpha''\alpha'''}\ldots a_{\alpha(p)\alpha}.$$

Dacă reluăm același procedeu plecînd de la o valoare $a_{\delta\delta'}$, care nu face parte din produsul precedent, obținem un nou produs

$$a_{\delta\delta'}a_{\delta'\delta''}a_{\delta''\delta'''}\ldots a_{\delta(q)\delta}$$

etc.

Să observăm acum că

$$a_{\alpha\alpha'}a_{\alpha'\alpha''}\ldots a_{\alpha(p)\alpha}=p_{\alpha\alpha'}p_{\alpha'\alpha''}\ldots p_{\alpha(p)\alpha}e^{(a_\alpha+a_{\alpha'}+\ldots+a_{\alpha(p)}-p\theta)\,it}=$$

$$=p_{\alpha\alpha'}p_{\alpha'\alpha''}\ldots p_{\alpha(p)\alpha}e^{p\,(H-\theta)\,it},$$

unde

$$\frac{a_\alpha+a_{\alpha'}+\ldots+a_{\alpha(p)}}{p}=H.$$

In virtutea lui (XX.10,6), vom avea de asemenea

$$\frac{a_\delta+a_{\delta'}+\ldots+a_{\delta'q)}}{q}=H;$$

rezultă că

$$B_{m-1}=\sum(\pm a_{\alpha\alpha'}a_{\alpha'\alpha''}\ldots a_{\mu\mu'})=\sum(\pm p_{\alpha\alpha'}p_{\alpha'\alpha''}\ldots p_{\mu\mu'})\,e^{(n-s)\,(H-\theta)\,it}=$$

$$=B_{m-s}(0)\,e^{(n-s)\,(H-\theta)\,it}.$$

Deci

$$D(\lambda;\,t;\,\theta)=\lambda^m+B_1(0)\,e^{(H-\theta)\,it}\lambda^{m-1}+B_2(0)\,e^{2\,(H-\theta)\,it}\lambda^{m-2}+\ldots$$

$$\ldots+B_s(0)\,e^{s\,(H-\theta)\,it}\lambda^{m-s}+\ldots+B_m(0)\,e^{m\,(H-\theta)\,it}. \qquad \text{(XX.10,9)}$$

Din relația (XX.10,8) rezultă imediat că

$$\lambda=e^{(H-\theta)\,it}$$

este o rădăcină a lui $D(\lambda;\,t;\,\theta)=0$, ceea ce demonstrează afirmația noastră.

Egalitatea (XX.10,6) exprimă deci condiția necesară și suficientă pentru ca să existe rădăcinile $\lambda(t;\ \theta)$ de modul egal cu unitatea. Dacă în ecuația (XX.10,9) punem

$$\lambda = e^{(H-\theta)\,it}\mu,$$

obținem

$$\mu^m + B_1(0)\,\mu^{m-1} + \ldots + B_s(0)\,\mu^{m-s} + \ldots + B_m(0) = 0,$$

identic cu ecuația

$$D(\mu;\ 0;\ \theta) = 0.$$

Rădăcinile de modul egal cu unitatea ale acestei ecuații verifică o ecuație binomă, deci rădăcinile lui $D(\lambda;\ t;\ \theta)=0$ au forma

$$\lambda_j = e^{(H-\theta)\,it}\mu_j,$$

unde μ_j verifică ecuația binomă $\mu^p - 1 = 0$.
Pentru a calcula valoarea expresiei

$$\frac{\sum_{h=1}^{s} G_{kh}[\lambda_r(0);\ 0]}{\Psi_r[\lambda_r(0)]},$$

adunăm la coloana k a determinantului $D(\lambda;\ 0;\ \theta)$ pe toate celelalte și dezvoltăm apoi determinantul după elementele acestei coloane. Atunci

$$D(\lambda;\ 0;\ \theta) = (\lambda - 1)\sum_{h=1}^{s} G_{kh}(\lambda;\ 0)$$

sau încă

$$\sum_{h=1}^{s} G_{kh}(\lambda;\ 0) = \frac{D(\lambda;\ 0;\ \theta)}{\lambda - 1}. \qquad \text{(XX.10,10)}$$

Pe de altă parte, știm că

$$\Psi_r(\lambda) = \frac{D(\lambda;\ 0;\ \theta)}{(\lambda - \lambda_r)^{m_r}},$$

unde m este ordinul de multiplicitate al rădăcinii λ_r. Deci

$$\Psi_r[\lambda\ (0)] \neq 0,$$

de unde

$$\frac{\sum_{h=1}^{s} G_{hk}[\lambda_r(0);\ 0]}{\Psi_r[\lambda_r(0)]} = 0 \qquad \text{(XX.10,11)}$$

pentru orice rădăcină $\lambda_r(0)$ diferită de unitate.

Pentru $r=1$, dacă presupunem $\lambda_r=1$ rădăcină simplă, din relaţiile precedente obţinem

$$\frac{\sum_{h=1}^{s} G_{hk}[\lambda_1(0);\ 0]}{\Psi_1[\lambda_1(0);\ 0]} = 1. \qquad \text{(XX.10,11')}$$

Vom vedea de altfel că singurul caz interesant este acela în care $\lambda_r(0)=1$ este o rădăcină simplă.

§ 11. Matrice decompozabile şi indecompozabile

Operaţiile liniare folosite în studiul unui lanţ Markov introduc un număr de noţiuni dintre care cele relative la comportarea lor asimptotică sînt cele mai utile. Aşa se întîmplă cu noţiunea de matrice decompozabilă.

Matricea $\|p_{ik}\|$ este *indecompozabilă* dacă, fiind daţi doi indici oarecare h şi l, putem găsi întotdeauna un şir de probabilităţi diferite de zero astfel încît

$$p_{hk}\neq 0,\quad p_{k_1k_2}\neq 0,\ \ldots,\ p_{k_{s-1}k_s}\neq 0,\quad p_{kl}\neq 0,$$

cu alte cuvinte, dacă este întotdeauna posibil să trecem de la valoarea a_h la valoarea a_l pentru orice a_h şi a_l din mulţimea de valori $a_1, a_2, \ldots, a_m$.

Dacă condiţia precedentă nu este verificată, matricea este *decompozabilă*. In acest ultim caz, o simplă permutare între linii şi coloane o aduce la forma

$$\left\| \begin{matrix} A & B \\ C & D \end{matrix} \right\|,$$

unde cel puţin una dintre matricele C şi B este nulă. Dacă $B=C=0$, matricea este numită *total decompozabilă*. Dacă numai

una dintre aceste matrice este nulă, matricea de probabilități este *parțial decompozabilă.*

Continuînd această descompunere, vom ajunge la forma

$$\left\| \begin{array}{cccccc} A_1 & 0 & 0 \ldots 0 & 0 & 0 \\ 0 & A_2 & 0 \ldots 0 & 0 & 0 \\ 0 & 0 & A_3 \ldots 0 & 0 & 0 \\ \cdot & \cdot & \cdot \cdot \cdot \cdot \cdot \cdot & \cdot & \cdot \\ 0 & 0 & 0 \ldots A_{s-1} & 0 & 0 \\ 0 & 0 & 0 \ldots 0 & A_s & 0 \\ B_1 & B_2 & B_3 \ldots B_{s-1} & B_s & A_{s+1} \end{array} \right\|, \qquad \text{(A)}$$

unde $A_1, A_2, A_3, \ldots, A_s, A_{s+1}$ sînt matrice indecompozabile, $B_1, B_2, \ldots, B_s$ matrice diferite de zero și 0 matricea nulă.

§ 12. Momentul de ordinul întîi $M_k^{(1)}(n;\ \theta)$

Vom nota momentul de ordinul r al variabilei $X_{(n)}$, în ipoteza că

$$X_1 = \begin{pmatrix} a_1, & a_2, \ldots, a_m \\ p_{k_1}, & p_{k_2}, \ldots, p_{k_m} \end{pmatrix},$$

cu $M_k^{(r)}(n;\ \theta)$.

Știm că

$$M_k^{(1)}(n;\ \theta) = -i \left[\frac{\partial F_k(n;\ t;\ \theta)}{\partial t} \right]_{t=0}$$

și că

$$M_k^{(2)}(n;\ \theta) = - \left[\frac{\partial^2 F_k(n;\ t;\ \theta)}{\partial t^2} \right]_{t=0}.$$

Vom examina cazul în care indicele k aparține unei valori făcînd parte dintr-o mulțime finală.

Fie $a_1, a_2, \ldots, a_k, \ldots, a_q$ mulțimea finală din care face parte a_k. Relația (II) devine

$$F_k(n;\ t;\ \theta)=\sum_{h=1}^{m}\sum_{j=1}^{s}\frac{1}{(m_j-1)!}D_\lambda^{m_j-1}\left[\frac{\lambda^n G_{kh}(\lambda;\ t)}{\Psi_j(\lambda)}\right]_{\lambda=\lambda_j}, \qquad \text{(XX.12,1)}$$

toate elementele referindu-se la determinantul $D_1(\lambda;\ t;\ \theta)$ și la rădăcinile ecuației $D_1(\lambda;\ t;\ \theta)=0$.

Fie $\lambda_1(0;\ \theta)=1,\ \lambda_2(0;\ \theta),\ldots,\lambda_g(0;\ \theta)$ rădăcinile simple de modul egal cu unitatea ale ecuației $D_1(\lambda;\ 0;\ \theta)=0$.

Dacă considerăm acum rădăcinile

$$\lambda_1(t;\ \theta),\ \lambda_2(t;\ \theta),\ldots,\lambda_g(t;\ \theta)$$

ale ecuației $D_1(\lambda;\ t;\ \theta)=0$, care pentru $t\to 0$ tind către precedentele, putem lua întotdeauna pe t suficient de mic pentru ca fiecare dintre ele să fie simplă.

Putem atunci separa în relația (XX.12,1) acest grup de rădăcini de celelalte și obținem

$$F_k(n;\ t;\ \theta)=\sum_{j=1}^{g}\lambda_j^n\frac{G_{hk}(\lambda_j;\ t)}{\Psi_j(\lambda_j)}+H(\lambda;\ n), \qquad \text{(XX.12,2)}$$

cu

$$H(\lambda;\ n)=\sum_{j=g+1}^{q}\sum_{h=1}^{q}\frac{1}{(m_j-1)!}D_\lambda^{m_j-1}\left[\frac{\lambda^n G_{kh}(\lambda;\ t)}{\Psi_j(\lambda)}\right]_{\lambda=\lambda_j}.$$

Vom avea astfel

$$M_k(n;\ \theta)=\Bigg|-in\sum_{j=1}^{q}\lambda_j^{n-1}\lambda_j\sum_{h=1}^{q}\frac{G_{kh}(\lambda_j;\ t)}{\Psi_j(\lambda_j)}-$$

$$-\sum_{j=1}^{q}\lambda_j^n\sum_{h=1}^{q}\frac{\partial}{\partial t}\left[\frac{G_{kh}(\lambda_j;\ t)}{\Psi_j(\lambda_j)}\right]\Bigg|_{t=0}-iG(\lambda;\ n), \qquad \text{(XX.12,3)}$$

unde

$$G(\lambda;\ n)=\left[\frac{\partial}{\partial t}H(\lambda;\ n)\right]_{t=0}.$$

Să observăm mai întîi că utilizînd relația (XX.10,11) pentru $j=1$, termenul din membrul al doilea are valoarea

$$-in\lambda_1(0;\ \theta),$$

unde

$$\lambda_1(0;\ \theta)=\left[\frac{\partial\lambda_1(t;\ \theta)}{\partial t}\right]_{t=0}.$$

Să punem, pe de altă parte,

$$iA_j=\left[\sum_{h=1}^{q}\frac{\partial}{\partial t}\cdot\frac{G_{kh}(\lambda;\ t)}{\Psi_j(\lambda)}\right]_{t=0}.$$

Expresia momentului de ordinul întîi va fi atunci de forma

$$M_k(n;\ \theta)=-in\lambda_1+\sum_{j=2}^{q}A_j\lambda_j^n+G(n), \qquad \text{(XX.12,4)}$$

unde $\lambda_j\ (j=1,\ 2,\ldots,\ g)$ verifică o ecuație binomă

$$x^p-1=0 \quad \text{și} \quad -iG(n;\ \theta)=G(n).$$

Expresia (XX.12,3) pentru momentul de ordinul întîi corespunde problemei pe care ne-o punem relativ la acest moment, deoarece ea ne dă forma precisă pentru $n\to\infty$.

Într-adevăr, toate rădăcinile lui λ care figurează în $H(\lambda;\ n)$, deci și în $G(\lambda;\ n)$, sînt în modul mai mici decît unitatea pentru orice t, deci

$$G(\lambda;\ n)\to 0 \quad \text{cînd} \quad n\to\infty.$$

Asimptotic se obține momentul $M_k(n;\ \theta)$ compus dintr-o parte care tinde spre infinit ca $-i\lambda n^t$ și dintr-o parte periodică

$$\sum_{j=2}^{q}A_j\lambda_j^n.$$

Pentru calculul lui $\lambda_1'(0;\ \theta)$ vom observa că ecuația $D_1(\lambda;\ t;\ \theta)=0$ mai poate fi scrisă

$$\begin{vmatrix} \lambda e^{(\theta-a_1)it}-p_{11} & -p_{12}\ldots & -p_{1q} \\ -p_{21} & \lambda e^{(\theta-a_2)it}-p_{22}\ldots & -p_{2q} \\ \cdots\cdots\cdots & \cdots\cdots\cdots & \cdots\cdots \\ -p_{q1} & -p_{q2}\ldots & \lambda e^{(\theta-a_q)it}-p_{qq} \end{vmatrix}=0.$$

Dacă facem aici $\lambda=\lambda_1$ şi dacă derivăm în raport cu t, făcînd apoi $t=0$, obţinem

$$\lambda_1' \sum_{h=1}^{q} P_h - i \sum_{h=1}^{q} (a_h-\theta) P_h = 0$$

sau

$$\lambda_1' = \frac{i \sum_{h=1}^{q} P_h (a_h-\theta)}{\sum_{h=1}^{q} P_h},$$

notînd cu P_h minorul lui $1-p_{hh}$ în determinantul precedent, după ce am făcut mai întîi $t=0$.
Obţinem, aşadar, forma

$$M_k(n;\theta) = \frac{\sum_{h=1}^{q} P_h (a_h-\theta)}{\sum_{h=1}^{q} P_h} n + \sum_{j=2}^{q} A_j \lambda_j^n + G(n), \quad \text{(XX.12,5)}$$

care pentru valoarea

$$\theta=\bar{\theta}=\frac{\sum_{h=1}^{q} P_h a_h}{\sum_{h=1}^{q} P_h}$$

se reduce la partea periodică

$$\sum_{j=2}^{q} A_j \lambda_j^n.$$

§ 13. Momentul de ordinul al doilea $M_k^{(2)}(n;\theta)$

Momentul de ordinul al doilea al variabilei $X_{(n)}$ va fi calculat cu ajutorul aceleiaşi expresii fundamentale (II) a funcţiei caracteristice, avînd însă grijă să facem $\theta=\bar{\theta}$ astfel ca momentul $M_k(n;\theta)$ să rămînă periodic cînd $n\to\infty$.

In cazul unei mulțimi finale, avem

$$\bar{\theta}=\frac{\sum_{h=1}^{q} P_h a_h}{\sum_{h=1}^{q} P_h}.$$

Cum

$$\left[\frac{\partial\lambda_1(t;\bar{\theta})}{\partial t}\right]_{t=0}=\lambda_1'(0,\bar{\theta})=\frac{\sum_{h=1}^{q} P_h(a_h-\theta)}{\sum_{h=1}^{q} P_h},$$

rezultă că

$$\lambda_1'(0,\bar{\theta})=0. \qquad \text{(XX.13,1)}$$

Această egalitate este valabilă pentru toate rădăcinile $\lambda_j(0,\bar{\theta})$ de modul 1.

Pentru a demonstra aceasta, să ne amintim că aceste rădăcini verifică ecuația binomă $x^p-1=0$ și să construim lanțul

$$\begin{aligned} Y_1&=X_1+X_2+\ldots+X_p,\\ Y_2&=X_{p+1}+X_{p+2}+\ldots+X_{2p} \qquad \text{(XX.13,2)}\\ &\cdots\cdots\cdots \end{aligned}$$

Un calcul simplu ne arată că între rădăcinile $\lambda_j(t;\theta)$ din primul lanț și acelea din al doilea, pe care le vom nota cu $\mu_j(t;\theta)$, există relația

$$\mu_j(t;\theta)=\lambda_j^p(t;\theta), \qquad \text{(XX.13,3)}$$

de unde, prin derivare,

$$\mu_j'(t;\theta)=p\lambda_j^{p-1}(t;\theta)\lambda_j'(t;\theta). \qquad \text{(XX.13,4)}$$

Dacă în relația de mai sus facem $\theta=\bar{\theta}$, vom avea $\mu_j'(0;\bar{\theta})=0$ ori de cîte ori $\mu_j(t;\bar{\theta})=1$, deoarece în acest caz momentul de ordinul întîi al expresiei $X_1+X_2+\ldots+X_{np}-np\bar{\theta}=Y_1+Y_2+\ldots+Y_n-np\theta$ nu mai este asimptotic de ordinul n. Dar $\mu_j(0;\bar{\theta})=0$ atrage după sine, ținînd seama de (XX.13,4), $\lambda_j'(0;\theta)=0$, cu condiția să avem $\mu_j(0;\bar{\theta})=\lambda_j^p(0;\bar{\theta})=1$, adică ceea ce voiam să demonstrăm.

Dacă reluăm ecuaţia diferenţială

$$\frac{\partial F_k(n;t;\bar{\theta})}{\partial t} = n\sum_{j=1}^{q}\lambda_j^{n-1}\lambda'\sum_{h=1}^{q}\frac{G_{hk}(\lambda_j;t)}{\psi_j(\lambda_j)} + \\ + \sum_{j=1}^{q}\lambda_j^n\sum_{h=1}^{q}\frac{\partial}{\partial t}\left[\frac{G_{hk}(\lambda;t)}{\psi_j(\lambda_j)}\right] + G(\lambda;n) \qquad \text{(XX.13,5)}$$

şi dacă derivăm încă o dată, făcînd apoi $t=0$ şi utilizînd relaţiile $\lambda_j'(0;\bar{\theta})=0$ $(j=1,2,\ldots,q)$ şi relaţiile (XX.10,11) şi (XX.10,11′), obţinem

$$\left[\frac{\partial^2 F_k}{\partial t^2}\right]_{t=0} = \lambda_1''(0;\bar{\theta})\,n - \sum_{j=1}^{q}B_j\lambda_j^n - H(n), \qquad \text{(XX.13,6)}$$

unde B_j sînt constante, λ_j' rădăcinile de modul unitate, $H(n)$ o expresie a cărei valoare asimptotică tinde către zero şi unde am pus

$$\lambda_1''(0;\bar{\theta}) = \left[\frac{\partial^2\lambda_1(t;\bar{\theta})}{\partial t^2}\right]_{t=0}.$$

Relaţia (XX.13,5) ne dă pentru momentul de ordinul al doilea expresia

$$M_k^{(2)}(n;\theta) = -\lambda_1''(0;\bar{\theta})\,n - \sum_{j=1}^{q}B_j^n + H(n). \qquad \text{(XX.13,7)}$$

Se obţine astfel o formă de independenţă asimptotică în sensul că *coeficientul* $-\lambda_1''(0;\bar{\theta})$ *al termenului principal este independent de* k. Condiţia $D_1(\lambda;t;\theta)=0$, verificată de $\lambda_1(t;\theta)$, o vom pune sub forma

$$\begin{vmatrix} \lambda_1 e^{(\bar{\theta}-a_1)it}-p_{11} & -p_{12} & \ldots & -p_{1q} \\ -p_{21} & \lambda_1 e^{(\bar{\theta}-a_2)it}-p_{22} & \ldots & -p_{2q} \\ \ldots & \ldots & \ldots & \ldots \\ -p_{q1} & -p_{q2} & \ldots & \lambda_1 e^{(\bar{\theta}-a_q)it}-p_{qq} \end{vmatrix} = 0.$$

Dacă derivăm de două ori și ținem seama că $\lambda_1(0;\bar{\theta})=1$, $\lambda_1'(0;\bar{\theta})=0$, obținem

$$\lambda_1''(0;\bar{\theta})=\frac{\sum_{r=1}^{q}P_r(a_r-\bar{\theta})^r+2\sum_{r,s=1}^{m}P_{rs}(a_r-\bar{\theta})(a_s-\bar{\theta})}{\sum_{r=1}^{m}P_r}, \qquad \text{(XX.13,8)}$$

unde P_{rs} este minorul determinantului precedent, obținut după suprimarea liniilor și coloanelor de ordinul r și s și după ce am făcut $t=0$.

§ 14. Legea limită

Avem

$$F_k(n;t;\bar{\theta})=1+\frac{M_k(n;\bar{\theta})}{1!}it+\frac{M_k^{(2)}(n;\bar{\theta})}{2!}i^2t^2+\ldots,$$

cu

$$M_k(n;\bar{\theta})=X_k(n),$$

$$M_k^{(2)}(n;\bar{\theta})=-n\lambda_1''(0;\bar{\theta})+\Phi_k(n),$$

unde $X_k(n)$ și $\Phi_k(n)$ sînt funcții asimptotic periodice de n.
Dacă considerăm variabila

$$X^{(n)}=\frac{X_1+X_2+\ldots+X_n-n\bar{\theta}}{\sqrt{M_k^{(2)}(n;\bar{\theta})}}=\frac{X_1+X_2+\ldots+X_n-n\bar{\theta}}{\sqrt{-n\lambda_1''(0;\bar{\theta})+\Phi_k(n)}},$$

vom avea

$$M_k(X^{(n)})=\frac{X_k^{(n)}}{\sqrt{-n\lambda_1''+\Phi_k^{(n)}}},$$

cînd $n\to\infty$, afară de cazul special cînd $\lambda_1''(0;\bar{\theta})=0$, caz pe care îl lăsăm deoparte pentru un moment.
Să reluăm formula care dă funcția caracteristică

$$F_k(n;t;\bar{\theta})=\sum_{j=1}^{g}\lambda_j^n(t;\bar{\theta})\sum_{h=1}^{q}\frac{G_{hk}(\lambda_j;t)}{\psi_j(\lambda_j)}+$$

$$+\sum_{h=g+1}^{m}\sum_{h=g+1}^{m}\frac{1}{(m_j-1)!}D_\lambda^{m_j-1}\left[\frac{\lambda^n G_{kh}(\lambda;t)}{\psi_j(\lambda_j)}\right]_{\lambda=\lambda_j}.$$

Să punem peste tot

$$t=\frac{u}{\sqrt{-\lambda_1'' n}}$$

și să facem pe n să tindă către infinit. Cum modulele lui λ_j din cel de al doilea termen sînt mai mici decît unitatea pentru orice t, ca și pentru $t=0$, acest termen tinde către zero cînd $n\to\infty$.
Pe de altă parte, știm că

$$\lim_{n\to\infty}\sum_{h=1}^{q}\frac{G_{hk}\left(\lambda_j\left(\frac{u}{\sqrt{-\lambda_1'' n}};\bar{\theta}\right);\frac{u}{\sqrt{-\lambda_1'' n}}\right)}{\psi_j\left(\lambda_j\left(\frac{u}{\sqrt{-\lambda_1'' n}}\right)\right)}=1 \text{ sau } 0,$$

după cum $\lambda_j(0;\bar{\theta})=1$ sau numai $|\lambda_j(0;\bar{\theta})|=1$. Vom avea atunci

$$\lim_{n\to\infty}F_k\left(n;\frac{u}{\sqrt{-\lambda_1'' n}};\bar{\theta}\right)=\lim_{n\to\infty}\lambda_1^n\left[\frac{u}{\sqrt{-\lambda_1'' n}};\bar{\theta}\right]=$$

$$=\lim_{n\to\infty}\left[1-\frac{u^2}{2n}+\ldots\right]^n=e^{-\frac{u^2}{2}},$$

convergența fiind uniformă.
Putem deci trage concluzia că *în cazul* (II) *legea limită redusă a lanțului este legea normală a lui Laplace.*

§ 15. Legile limită pentru sumele de vectori înlănțuiți

Să considerăm într-un spațiu euclidian $s(>1)$-dimensional, m vectori $\vec{v}^1, \vec{v}^2, \ldots, \vec{v}^m$, avînd respectiv componentele

$$\vec{v}^h=(v_1^h, v_2^h, \ldots, v_s^h) \qquad (h=1, 2, \ldots, m).$$

Să considerăm vectorul aleator $\vec{V}(n)$ care la momentul n poate lua una dintre valorile $\vec{v}^h$, după regula dată de un lanț Markov discret, simplu și omogen

$$P[\vec{V}(n)=\vec{v}^k/V(n+1)=v^h]=p_{hk}.$$

Atunci, evident

$$p_{hk} \geqslant 0, \quad \sum_{k=1}^{m} p_{hk} = 1.$$

Să considerăm vectorul

$$\vec{W}(n) = \vec{V}(1) + \vec{V}(2) + \ldots + \vec{V}(n).$$

Ne propunem să determinăm comportarea asimptotică a vectorului aleator $\vec{W}(n)$ dacă $n \to \infty$. Vom presupune

$$\vec{W}(n) = (W_1(n),\ W_2(n), \ldots,\ W_s(n)).$$

Să notăm cu $P_k(\lambda_1,\ \lambda_2, \ldots, \lambda_m;\ n)$ probabilitatea evenimentului

$$\vec{W}(n) = \lambda_1 \vec{v}^1 + \lambda_2 \vec{v}^2 + \ldots + \lambda_m \vec{v}^m,$$

știind că

$$P(\vec{W}(1) = \vec{v}^k) = p_{hk} \quad (h = 1,\ 2, \ldots, m).$$

Atunci relația de recurență devine

$$P_k(\lambda_1,\ \lambda_2, \ldots, \lambda_m;\ n) =$$

$$= \sum_{h=1}^{m} p_{hk} P_h(\lambda_1,\ \lambda_2, \ldots, \lambda_{h-1},\ \lambda_h - 1,\ \lambda_{h+1}, \ldots, \lambda_m;\ n-1). \qquad \text{(XX.15,1)}$$

Să considerăm vectorul

$$\vec{\theta} = (\theta_1,\ \theta_2, \ldots, \theta_s),$$

ale cărui componente nu sînt deocamdată determinate.

Funcția caracteristică a variabilei $\vec{W}(t) - n\vec{\theta}$ este, prin definiție,

$$F_k(n;\ \vec{t};\ \vec{\theta}) = \sum_{\lambda_1, \lambda_2, \ldots, \lambda_m} P_k(\lambda_1,\ \lambda_2, \ldots, \lambda_m;\ n)\, e^{\lambda_1 (\vec{v}^1, \vec{t})\, i + \ldots + \lambda_m (\vec{v}^m, \vec{t})\, i - n (\vec{\theta}, \vec{t})\, i},$$

unde $\vec{t} = (t_1,\ t_2, \ldots, t_s)$, iar $(\vec{a},\ \vec{b})$ este produsul scalar al vectorilor $\vec{a}$ și $\vec{b}$.

Ținînd seama de (XX.15,1), din relația precedentă obținem

$$F_k(n;\vec{t};\vec{\theta})=\sum_{j=1}^{m} p_{kj}e^{(\vec{\omega j},\vec{t})i-(\vec{\theta},\vec{t})i}F_j(n-1;\vec{t};\vec{\theta}).$$

Această relație liniară ne va servi pentru rezolvarea problemei.
Să introducem determinantul

$$A=\|p_{kj}e^{(\vec{\omega j},\vec{t})i-(\vec{\theta},\vec{t})i}\|$$

și fie E matricea unitară cu m linii și m coloane. Să notăm cu $\lambda_i\,(i=1,2,\ldots,r)$ rădăcinile distincte de ordinele de multiplicitate respective $m_i\,(i=1,2,\ldots,r)$ ale ecuației

$$D(\lambda;\vec{t};\vec{\theta})=\|A-\lambda E\|=0.$$

Atunci

$$\|A-\lambda E\|=(\lambda-\lambda_1)^{m_1}(\lambda-\lambda_2)^{m_2}\ldots(\lambda-\lambda_r)^{m_r}.$$

Să considerăm funcția

$$\psi_j(\lambda;\vec{t};\vec{\theta})=\frac{\|A-\lambda E\|}{(\lambda-\lambda_j)^{m_j}}.$$

Fie de asemenea $G_{hk}(\lambda;\vec{t};\vec{\theta})$ minorul elementului din linia h și coloana k din determinantul $\|A-\lambda E\|$.
Cu notațiile de mai sus, obținem

$$F_k(n;\vec{t};\vec{\theta})=\sum_{k=1}^{m}\sum_{j=1}^{r}D_\lambda^{m_j-1}\left[\frac{\lambda^n G_{hk}(\lambda;\vec{t};\vec{\theta})}{\psi_j(\lambda)}\right]_{\lambda=\lambda_j}, \qquad \text{(XX.15,2)}$$

unde D_λ^p reprezintă derivata de ordinul p în raport cu λ.

Fie $\lambda_j(\vec{t};\vec{\theta})$ o rădăcină a ecuației $D(\lambda;\vec{t};\vec{\theta})=0$.

Teorema 1. *Intotdeauna*

$$|\lambda_j(\vec{t};\vec{\theta})|\leqslant 1, \qquad \text{(XX.15,3)}$$

egalitatea avînd loc pentru orice $\vec{t}$, *numai dacă pentru*

$$p_{j_1 j_2}>0,\ p_{j_1 j_3}>0,\ \ldots,\ p_{j_\sigma j_1}>0$$

rezultă

$$\frac{v_h^{j_1}+v_h^{j_2}+\ldots+v_h^{j_\sigma}}{\sigma}=b_h, \qquad \text{(XX.15,3')}$$

pentru orice $h=1,\ 2,\ \ldots,\ s$.

Demonstrația acestei teoreme rezultă folosind aceeași cale ca în cazul lanțului obișnuit.

Pentru determinarea cazului de egalitate, observăm că putem găsi întotdeauna m valori z_i, astfel încît

$$\sum_{h=1}^{m} p_{jh} z_h=\lambda_j(\vec{t};\ \vec{\theta})\, e^{-(\vec{v^j},\ \vec{u})\, i+(\vec{v},\ \vec{t})\, i} z_j. \qquad \text{(XX.15,4)}$$

Pentru ca inegalitatea (XX.15,3) să devină egalitate ($j=1,\ 2,\ldots,m$), se vede imediat că numerele z_h trebuie să aibă același modul, pentru orice h. Putem deci nota

$$z_h=\rho e^{i\omega_h(\vec{t},\ \vec{\theta})},$$

$$\lambda_g(\vec{t};\ \vec{\theta})=e^{i\varphi_g(\vec{t},\ \vec{\theta})}.$$

Cu aceste notații, relația devine

$$\sum_{j=1}^{m} p_{kj} e^{i\omega_k(\vec{t};\ \vec{\theta})-i\omega_h(\vec{t};\ \vec{\theta})-i\varphi_g(\vec{t};\ \vec{\theta})+(\vec{v^j},\ \vec{t})\, i-(\vec{\theta},\ \vec{t})\, i}=1.$$

Rezultă că

$$\omega_k(\vec{t};\ \vec{\theta})-\omega_h(\vec{t};\ \vec{\theta})-\varphi_g(\vec{t};\ \vec{\theta})+(\vec{v^j},\ \vec{t})-(\vec{\theta},\ \vec{t})=2k\pi,$$

dacă $p_{kj}>0$, k' fiind un număr întreg, pozitiv sau negativ.

Fie

$$p_{h_1 h_2}>0,\quad p_{h_2 h_3}>0,\ \ldots,\ p_{h_{\sigma-1} h_\sigma}>0,\quad p_{h_\sigma h_1}>0.$$

Adunînd relațiile corespunzătoare șirului de mai sus, rezultă imediat

$$\varphi_g(\vec{t};\ \vec{\theta}) = \frac{(\vec{v}^{h_1},\ \vec{t})+\ldots+(\vec{v}^{h_\sigma},\ \vec{t})}{\sigma} - (\vec{\theta},\ \vec{t}) + \frac{2k\pi}{\sigma}.$$

Pentru un alt șir

$$p_{i_1 i_2}>0,\quad p_{j_2 j_3}>0,\ \ldots,\ p_{j_{\sigma-1} j_\sigma}>0,\quad p_{i_\sigma j_1}>0,$$

deducem în mod analog

$$\varphi_g(\vec{t};\ \vec{\theta}) = \frac{(\vec{v}^{j_1},\ \vec{t})+\ldots+(\vec{v}^{j_\rho},\ t)}{\rho} - (\vec{\theta},\ \vec{t}) + \frac{2k\pi}{\rho}$$

sau

$$\frac{(\vec{v}^{h_1},\ \vec{t})+\ldots+(\vec{v}^{h_\sigma},\ \vec{t})}{\sigma} = \frac{(\vec{v}^{j_1},\ \vec{t})+\ldots+(\vec{v}^{j_\sigma},\ \vec{t})}{\rho}.$$

Condițiile de mai sus sînt deci necesare.

Ecuația

$$D(\lambda;\ \vec{0};\ \vec{\theta}) = 0$$

admite întotdeauna rădăcina $\lambda = 1$. Presupunem că această rădăcină este simplă.

Fie $\lambda_1(\vec{t};\ \vec{\theta})$ această rădăcină,

$$\lambda_1(\vec{0},\ \vec{\theta}) = 1.$$

Se demonstrează în același fel că

$$\sum_{h=1}^{m} \frac{G_{kh}(\lambda_r(\vec{t};\ \vec{\theta});\ \vec{t})}{\psi_r(\lambda_r(\vec{t};\ \vec{\theta});\ \vec{t})} = 0, \quad \text{dacă} \quad u = 0\ (r \neq 1);$$

$$\sum_{h=1}^{m} \frac{G_{kh}(\lambda_1(\vec{t};\ \vec{\theta});\ \vec{t})}{\psi_1(\lambda_1(\vec{t};\ \vec{\theta});\ \vec{t})} = 1, \quad \text{dacă} \quad u = 0.$$

Fie

$$\lambda_1(\vec{0};\ \vec{\theta}) = 1,\ \ \lambda_2(\vec{0};\ \vec{\theta}),\ \ldots,\ \lambda_g(\vec{0};\ \vec{\theta})$$

rădăcinile de modul unitate ale ecuației $D(\lambda; \vec{0}; \vec{\theta})$. Putem lua pe t suficient de mic, pentru ca rădăcinile

$$\lambda_1(\vec{t}; \vec{\theta}), \quad \lambda_2(\vec{t}; \vec{\theta}), \ldots, \lambda_g(\vec{t}; \vec{\theta})$$

ale ecuației $D(\lambda; \vec{t}; \vec{\theta})$ să fie de asemenea simple.
Putem scrie în acest caz

$$F_k(n; \vec{t}; \vec{\theta}) = \sum_{j=1}^{q} \lambda_j^n(\vec{t}; \vec{\theta}) \sum_{h=1}^{m} \frac{G_{kh}(\lambda_j; \vec{t})}{\psi_j(\lambda_j)} + H(\lambda; n) \quad \text{(XX.15,5)}$$

unde

$$H(\lambda; n) = \sum_{j=1}^{m} \sum_{h=1}^{q} \frac{1}{(m_j-1)!} D_\lambda^{m_j-1} \left[\frac{\lambda^n G_{kh}(\lambda; \vec{t})}{\psi_j(\lambda)} \right|_{\lambda=\lambda_j}.$$

Modulul tuturor rădăcinilor λ_j care figurează în $H(\lambda; n)$ este mai mic decît unitatea, astfel că

$$\lim_{n\to\infty} H(\lambda; n) = 0 \qquad \text{(XX.15,6)}$$

pentru orice $\vec{t}$.
Derivînd relația (XX.15,5) și ținînd seama de (XX.15,6), rezultă

$$M[W_k(n) - n\theta_k] = in \left[\frac{\partial \lambda_1(\vec{t}; \vec{\theta})}{\partial t_k} \right]_{t=0} + \sum_{j=2}^{g} A_j \lambda_j^n(\vec{0}; \vec{\theta}) + G(n)$$

$$(k = 1, 2, \ldots, m; \; j = 2, 3, \ldots, g),$$

unde A_j sînt niște constante $\lambda_j(\vec{0}; \vec{\theta})$, care verifică ecuația binomă $\lambda^g - 1 = 0$, iar $G(n) \to 0$, pentru $n \to \infty$.

Putem acum determina pe $\vec{\theta}$, astfel încît

$$\left[\frac{\partial \lambda_1(\vec{t}; \vec{\theta})}{\partial t_k} \right]_{t=0} = 0 \quad (k = 1, 2, \ldots, m).$$

Scanned with OKEN Scanner

Să considerăm determinantul fundamental

$$\Delta=\begin{vmatrix} p_{11}-1 & p_{12} & \dots p_{1m} \\ p_{21} & p_{22}-1 \dots & p_{2m} \\ \dots & \dots & \dots \\ p_{m1} & p_{m2} & \dots p_{mm}-1 \end{vmatrix}$$

şi să notăm cu P_h minorul elementului $p_{hh}-1$ din Δ.

Ţinînd seamă că $\lambda_1(\vec{0};\ \vec{\theta})=1$, prin derivarea egalităţii $D(\lambda;\ \vec{t};\ \vec{\theta})=0$ se ajunge la rezultatul că trebuie să avem

$$\theta_k=\frac{\sum_{h=1}^{m} P_h v_k^h}{\sum_{h=1}^{m} P_h}. \qquad \text{(XX.15,7)}$$

De aici înainte, vom considera pentru $\vec{\theta}=(\theta_1, \theta_2, \dots, \theta_s)$ valorile de mai sus.

Derivînd din nou, obţinem

$$\left[\frac{\partial^2\lambda_1(\vec{t};\ \vec{\theta})}{\partial t_h \partial t_k}\right]_{\vec{t}=0}=\frac{\sum_{j=1}^{m}(v_h^j-\theta_h)(v_k^j-\theta_k)P_j-\sum_{i,\,j=1}^{m}(v_h^i-\theta_h)(v_k^j-\theta_k)P_{ij}}{\sum_{j=1}^{m}P_j},$$

unde P_{ij} este determinantul obţinut din Δ prin suprimarea liniilor şi coloanelor de rangul i şi j.

Pe de altă parte, derivînd de două ori funcţia caracteristică (XX.15,5), obţinem

$$M[W_k(n)-n\theta_k]^2=n\left[\frac{\partial^2\lambda_1(\vec{t};\ \vec{\theta})}{\partial t_k^2}\right]_{\vec{t}=0}+O(1),$$

$$M[(W_h(n)-n\theta_h)(W_k(n)-n\theta_k)]=-n\left[\frac{\partial^2\lambda_1(\vec{t};\ \vec{\theta})}{\partial t_h\,\partial t_k}\right]_{\vec{t}=0}+O(1).$$

Am obţinut astfel comportarea asimptotică a valorilor medii

$$M[W_k(n)-n\theta_k]^2 \quad \text{şi} \quad M[(W_h(n)-n\theta_h)(W_k(n)-n\theta_k)].$$

Să notăm acum cu $f_k(n;\vec{t})$ funcția caracteristică a variabilei

$$\frac{\vec{W}(n)-n\vec{\theta}}{\sqrt{n}}.$$

Atunci

$$f_k(n;\vec{t})=F_k\left(n;\frac{\vec{t}}{\sqrt{n}};\vec{\theta}\right).$$

Ținînd seama de rezultatele obținute, avem

$$\lim_{n\to\infty} f_k(n;\vec{t})=\lim_{n\to\infty}\lambda_1^n\left(\frac{\vec{t}}{\sqrt{n}};\vec{\theta}\right).$$

Dezvoltînd pe λ_1 în raport cu $\vec{t}$, obținem

$$\ln\lambda_1\left(\frac{\vec{t}}{\sqrt{n}};\vec{\theta}\right)=-\frac{1}{2n}\left[\sum_{h,k=1}^{m}\left(\frac{\partial^2\lambda_1(\vec{t};\vec{\theta}}{\partial t_h\partial t_k}\right)_{\vec{t}=0}t_h t_k\right]+O\left(n^{-\frac{1}{2}}\right),$$

pentru orice $|\vec{t}|<T$, de unde rezultă

$$\lim_{n\to\infty} f_k(n;\vec{t})=\exp\left\{-\frac{1}{2}\sum_{h,k=1}^{m}\left(\frac{\partial^2\lambda_1(\vec{t};\vec{\theta})}{\partial t_h\,\partial t_k}\right)_{\vec{t}=0}t_h t_k\right\}.$$

Să introducem variabilele

$$\xi_i=\sum_{j=1}^{s}(v_h^i-a_h)\,t_h;$$

rezultă atunci că

$$\lim_{n\to\infty} f_k(n;\vec{t})=\exp\left\{\frac{\frac{1}{2}\sum_{i=1}^{s}P_i\xi_i^2-2\sum_{i,j=1}P_{ij}\xi_i\xi_j}{\sum_{j=1}^{s}P_j}\right\}.$$

Putem deci enunța

Teorema 2. *Dacă probabilităţile* p_{hk} *corespund unei mulţimi finale, vectorul*

$$\frac{\vec{W}(n)-n\vec{\theta}}{\sqrt{n}}$$

urmează, la limită, legea normală multidimensională, a cărei funcţie caracteristică este

$$e^{\frac{1}{2}\sum_{i=1}^{s} p_i\xi_i^2-2\sum_{i,j=1}^{s} p_{ij}\xi_i\xi_j \Big/ \sum_{j=1}^{s} p_j} . \qquad \text{(XX.15,8)}$$

§ 16. Funcţii de estimaţie eficiente pentru lanţuri Markov

Să considerăm o selecţie de mărime n

$$x_1, x_2, \ldots, x_n$$

formată din n variabile dependente, care se succed după regula unui lanţ Markov simplu şi constant.

Fie

$$f(x, y; \alpha)\, dy$$

probabilitatea ca să avem $y<x_p<y+dy$, condiţionată de faptul că $x_{p-1}=x$. Această probabilitate, după cum ştim, rămîne aceeaşi oricare ar fi fost valorile luate de variabilele precedente $x_{p-2}, x_{p-3}, \ldots$ $\ldots, x_2, x_1$. α este un parametru, a cărui valoare necunoscută trebuie s-o determinăm pe baza selecţiei.

Este evident că trebuie să avem

$$f(x, y; \alpha) \geqq 0, \quad \int_{-\infty}^{+\infty} f(x, y; \alpha)\, dy = 1, \qquad \text{(XX.16,1)}$$

oricare ar fi x şi α.

Fie $\bar{\alpha}(x_1, x_2, \ldots, x_n)$ o funcţie de selecţie. Intocmai ca la o selecţie cu variabile independente, vom spune că $\bar{\alpha}(x_1, x_2, \ldots, x_n)$ este o

funcție de estimație corectă a lui α, dacă pentru $n \to \infty$, $\overline{\alpha}(x_1, x_2, \ldots, x_n)$ tinde în probabilitate către α.

Dacă

$$M[\overline{\alpha}(x_1, x_2, \ldots, x_n)] = \alpha, \quad \underset{n\to\infty}{D^2}(\overline{\alpha}) \to 0,$$

$\overline{\alpha}(x_1, x_2, \ldots, x_n)$ este o funcție de estimație absolut corectă a lui α.

Se poate demonstra, întocmai ca și pentru selecțiile cu variabile independente, că pentru toate funcțiile de estimație absolut corecte există un minim al dispersiei. Funcția de estimație a cărei dispersie atinge acest minim poartă numele de *funcție de estimație eficientă*. Demonstrația, după cum se va vedea, nu diferă esențial de demonstrația dată în cazul variabilelor independente.

Admitem că

$$f(x, y; \alpha) > 0$$

și că $f(x, y; \alpha)$ este continuă în raport cu x, y, α, admițînd derivate de ordinul întîi în raport cu α.

Fie $f(x_0, y; \alpha)\, dy$ probabilitatea ca să avem $y < x_1 < y + dy$. In aceste condiții, ne propunem să determinăm o margine inferioară, diferită de zero, pentru dispersia funcțiilor de estimație absolut corecte.

Din (XX.16,1), derivînd în raport cu α, obținem

$$\int_{-\infty}^{+\infty} \frac{\partial f(x, y; \alpha)}{\partial \alpha}\, dy = 0$$

sau

$$M\left[\frac{\partial \ln f(x, y; \alpha)}{\partial \alpha}\right] = 0. \qquad \text{(XX.16,2)}$$

In această relație, α este variabila aleatorie, x și y fiind doi parametri.

Prin definiție

$$M[\overline{\alpha}(x_1, x_2, \ldots, x_n)] =$$

$$= \int\int \ldots \int_{R_n} \overline{\alpha}(\xi_1, \xi_2, \ldots, \xi_n)\, f(x_0, \xi_1; \alpha)\, f(\xi_1, \xi_2; \alpha) \ldots$$

$$\ldots f(\xi_{n-1}, \xi_n; \alpha)\, d\xi_1 \ldots d\xi_n, \qquad \text{(XX.16,3)}$$

integrala fiind luată pe întreg spațiul R_n al selecției.

Scanned with OKEN Scanner

Deoarece valoarea medie a lui $\overline{\alpha}(x_1, x_2, \ldots, x_n)$ este α, avem relația

$$\int\int \ldots \int_{R_n} \overline{\alpha}(\xi_1, \xi_2, \ldots, \xi_n) f(x_0, \xi_1; \alpha) f(\xi_1, \xi_2; \alpha) \ldots$$

$$\ldots f(\xi_{n-1}, \xi_n; \alpha) d\xi_1 \ldots d\xi_n = \alpha. \qquad \text{(XX.16,4)}$$

Derivînd din nou în raport cu α

$$\int\int \ldots \int_{R_n} \overline{\alpha}(\xi_1, \xi_2, \ldots, \xi_n) \sum_{i=1}^{n} \frac{\partial \ln f(\xi_{i-1}, \xi_i; \alpha)}{\partial \alpha} \prod_{i=1}^{n} f(\xi_{i-1}, \xi_i; \alpha) d\xi_i = 1$$

$$(\xi_0 = x_0). \qquad \text{(XX.16.5)}$$

Pe de altă parte, din (XX.16,2) rezultă

$$\alpha \int\int \ldots \int_{R_n} \sum_{i=1}^{n} \frac{\partial \ln f(\xi_{i-1}, \xi_i; \alpha)}{\partial \alpha} \prod_{i=1}^{n} f(\xi_{i-1}, \xi_i; \alpha) d\xi_i = 0 \quad (\xi_0 = x_0).$$

Prin urmare

$$\int\int \ldots \int_{R_n} [\overline{\alpha}(\xi_1, \xi_2, \ldots, \xi_n) - \alpha] \sum_{i=1}^{n} \frac{\partial \ln f(\xi_{i-1}, \xi_i; \alpha)}{\partial \alpha} \prod_{i=1}^{n} f(\xi_{i-1}, \xi_i; \alpha) d\xi_i = 1$$

$$(\xi_0 = x_0) \qquad \text{(XX.16,6)}$$

sau

$$M\left[(\overline{\alpha}(x_1, x_2, \ldots, x_n) - \alpha) \sum_{i=1}^{n} \frac{\partial \ln f(x_{i-1}, x_i; \alpha)}{\partial \alpha}\right] = 1. \qquad \text{(XX.16,6')}$$

Aplicînd inegalitatea lui Schwarz, obținem astfel

$$M[\overline{\alpha}(x_1, x_2, \ldots, x_n) - \alpha]^2 \cdot M\left[\sum_{i=1}^{n} \frac{\partial \ln f(x_{i-1}, x_i; \alpha)}{\partial \alpha}\right]^2 \geqq 1$$

sau

$$D^2[\overline{\alpha}(x_1, x_2, \ldots, x_n)] \geqq \frac{1}{M\left[\sum_{i=1}^{n} \frac{\partial \ln f(x_{i-1}, x_i; \alpha)}{\partial \alpha}\right]^2}. \qquad \text{(XX.16,7)}$$

Membrul al doilea al acestei inegalități nu depinde de $\overline{\alpha}$. Prin urmare, proprietatea de minim pentru dispersia funcțiilor de estimație absolut corecte este demonstrată. Putem însă să punem expresia acestui minim sub o formă mai simplă și să vedem în ce caz există o funcție de estimație eficientă pentru α.

Avem

$$M\left[\sum_{i=1}^{n}\frac{\partial \ln f(x_{i-1}, x_i; \alpha)}{\partial\alpha}\right]^2 =$$

$$= \int\int\ldots\int_{R_n}\left[\sum_{i=1}^{n}\frac{\partial \ln f(x_{i-1}, x_i; \alpha)}{\partial\alpha}\right]^2 \prod_{i=1}^{n} f(\xi_{i-1}, \xi_i; \alpha)\, d\xi_i, \quad (\xi_0 = x_0).$$

Să observăm însă că

$$\int\int\ldots\int_{R_n}\frac{\partial \ln f(\xi_{i-1}, \xi_i; \alpha)}{\partial\alpha}\cdot\frac{\partial \ln f(\xi_{j-1}, \xi_j; \alpha)}{\partial\alpha}\prod_{i=1}^{n} f(\xi_{i-1}, \xi_i; \alpha)\, d\xi_i = 0$$

$$(\xi_0 = x_0)\ (i \neq j),$$

în baza egalității (XX.16,2).

Ca atare, rezultă

$$M\left[\sum_{i=1}^{n}\frac{\partial \ln f(x_{i-1}, x_i; \alpha)}{\partial\alpha}\right]^2 =$$

$$= \int\int\ldots\int_{R_n}\sum_{i=1}^{n}\left[\frac{\partial \ln f(\xi_{i-1}, \xi_i; \alpha)}{\partial\alpha}\right]^2 \prod_{i=1}^{n} f(\xi_{i-1}, \xi_i; \alpha)\, d\xi_i \quad \text{(XX.16,8)}$$

sau

$$M\left[\sum_{i=1}^{n}\frac{\partial \ln f(x_{i-1}, x_i; \alpha)}{\partial\alpha}\right]^2 = \sum_{i=1}^{n}\left(\frac{\partial \ln f(x_{i-1}, x_i; \alpha)}{\partial\alpha}\right)^2. \quad \text{(XX.16,8')}$$

Să notăm prin $F_i(x)\,dx$ probabilitatea $P(x < x_i < x + dx)$ $(i = 1, 2, \ldots, n)$.

Avem

$$\int\int\dots\int_{R_n}\left[\frac{\partial \ln f(\xi_{i-1}, \xi_i; \alpha)}{\partial\alpha}\right]^2 \prod_{i=1}^{n} f(\xi_{i-1}, \xi_i; \alpha)\, d\xi_i =$$

$$= \int\int_{R_2}\left(\frac{\partial \ln f(x, y; \alpha)}{\partial\alpha}\right) F_i(x) f(x, y; \alpha)\, dxdy,$$

ultima integrală fiind extinsă la tot planul.
Obţinem astfel

$$M\left(\sum_{i=1}^{n} \frac{\partial \ln f(x_{i-1}, x_i; \alpha)}{\partial\alpha}\right)^2 =$$

$$= \int\int_{R_2}\left(\frac{\partial \ln f(x, y; \alpha)}{\partial\alpha}\right)^2 \sum_{i=1}^{n} F_i(x) f(x, y; \alpha)\, dxdy, \quad \text{(XX.16,9)}$$

unde

$$F_i(x) = \int\int\dots\int_{R_{p-1}} f(x_0, \xi_1; \alpha) f(\xi_1, \xi_2; \alpha) \dots f(\xi_{p-1}, x; \alpha)\, d\xi_1\, d\xi_2 \dots d\xi_{p-1}. \quad \text{(XX.16,10)}$$

Inegalitatea (XX.16,7) se transformă în egalitate, ţinînd seamă de (XX.16,6), dacă

$$K(\alpha)[\overline{\alpha}(\xi_1, \xi_2, \dots, \xi_n) - \alpha] = \sum_{i=1}^{n} \frac{\partial \ln f(\xi_{i-1}, \xi_i; \alpha)}{\partial\alpha} \quad (\xi_0 = x_0), \quad \text{(XX.16.11)}$$

oricare ar fi $\xi_1, \xi_2, \dots, \xi_n$, prin $K(\alpha)$ notînd o funcţie oarecare de α.
In relaţia (XX.16,11) să punem

$$\xi_1 = x,\ \xi_2 = y,\ \xi_3 = 0, \dots,\ \xi_n = 0.$$

Rezultă că trebuie să avem, oricare ar fi x şi y,

$$K(\alpha)[\overline{\alpha}(x, y, 0, 0, \dots, 0) - \alpha] = \frac{\partial \ln f(x_0, x; \alpha)}{\partial\alpha} + \frac{\partial \ln f(x, y; \alpha)}{\partial\alpha} +$$

$$+ \frac{\partial \ln f(y, 0; \alpha)}{\partial\alpha} + (n-3)\frac{\partial \ln f(0, 0; \alpha)}{\partial\alpha}. \quad \text{(XX.16,12)}$$

Tot în relația (XX.16,11) să punem

$$\xi_1=0,\ \xi_2=x,\ \xi_3=y,\ \xi_4=0,\ldots,\ \xi_n=0;$$

obținem

$$K(\alpha)\,[\overline{\alpha}\,(0,\ x,\ y,\ 0,\ldots,\ 0)-\alpha]=\frac{\partial \ln f(x_0,\ 0;\ \alpha)}{\partial\alpha}+$$

$$+\frac{\partial \ln f(0,\ x;\ \alpha)}{\partial\alpha}+\frac{\partial \ln f(x,\ y;\ \alpha)}{\partial\alpha}+\frac{\partial \ln f(y,\ 0;\ \alpha)}{\partial\alpha}+$$

$$+(n-4)\frac{\partial \ln f(0,\ 0;\ \alpha)}{\partial\alpha}. \qquad \text{(XX.16,13)}$$

Scăzînd egalitatea (XX.16,11) din (XX.16,13) rezultă

$$K(\alpha)\,[\overline{\alpha}\,(0,\ x,\ y,\ 0,\ 0,\ldots,\ 0)-\overline{\alpha}\,(x,\ y,\ 0,\ldots,\ 0)]=$$

$$=\frac{\partial \ln f(0,\ x;\ \alpha)}{\partial\alpha}-\frac{\partial \ln f(0,\ 0;\ \alpha)}{\partial\alpha}.$$

Membrul întîi este funcție de x și y, pe cînd al doilea este funcție numai de x. Trebuie deci să avem

$$\overline{\alpha}\,(0,\ x,\ y,\ 0,\ 0,\ldots,\ 0)-\overline{\alpha}\,(x,\ y,\ 0,\ 0,\ \ldots,\ 0)=H(x),$$

prin $H(x)$ notînd o funcție de x.
Obținem astfel

$$\frac{\partial \ln f(0,\ x;\ \alpha)}{\partial\alpha}=K(\alpha)\,H(x)+\frac{\partial \ln f(0,\ 0;\ \alpha)}{\partial\alpha}. \qquad \text{(XX.16,14)}$$

In mod analog, punînd

$$\xi_1=0,\ \xi_2=0,\ldots,\ \xi_{n-2}=0,\ \xi_{n-1}=x,\ \xi_n=y,$$

și repetînd același raționament, obținem

$$\frac{\partial \ln f(y,\ 0;\ \alpha)}{\partial\alpha}=K(\alpha)\,\overline{H}(y)+\frac{\partial \ln f(0,\ 0;\ \alpha)}{\partial\alpha}. \qquad \text{(XX.16,15)}$$

Inlocuind aceste valori în (XX.16,13), rezultă

$$\frac{\partial \ln f(x,\ y;\ \alpha)}{\partial\alpha}=K(\alpha)\,[M(x,\ y)-\alpha]+G(\alpha), \qquad \text{(XX.16,16)}$$

unde $M(x,\ y)$ este o funcție de x și y.

Inlocuind această valoare în (XX.16,11), obţinem

$$K(\alpha)\cdot\bar{\alpha}(\xi_1, \xi_2, \ldots, \xi_n) - K(\alpha)\cdot\alpha =$$

$$= K(\alpha)\sum_{i=1}^{n} M(\xi_{i-1}, \xi_i) - n\alpha K(\alpha) + nG(\alpha).$$

De unde rezultă că, în afară de o constantă, care nu interesează, trebuie să avem

$$G(\alpha) = \frac{n-1}{n}\alpha K(\alpha).$$

Inlocuind această valoare în (XX.16,16), rezultă

$$\frac{\partial \ln f(x, y; \alpha)}{\partial \alpha} = K(\alpha)\left[M(x, y) - \frac{\alpha}{n}\right].$$

Această relaţie poate fi scrisă sub forma

$$\frac{\partial \ln f(x, y; \alpha)}{\partial \alpha} = L(\alpha)[A(x, y) - \alpha], \qquad \text{(XX.16,17)}$$

unde $L(\alpha)$ este o funcţie oarecare de α şi $A(x, y)$ o funcţie de x şi y.

Inlocuind această valoare în (XX.16,11), se vede că

$$\bar{\alpha}(x_1, x_2, \ldots, x_n) =$$

$$= \frac{A(x_0, x_1) + A(x_1, x_2) + \ldots + A(x_{n-1}, x_n)}{n}. \qquad \text{(XX.16,18)}$$

Deci, pentru ca să existe o funcţie de estimaţie eficientă, este necesar să avem relaţiile (XX.16,17) şi (XX.16,18).

Vom arăta că aceste relaţii sînt şi suficiente.

Din relaţia (XX.16,17) deducem

$$\ln f(x, y; \alpha) = \varphi'(\alpha)[A(x, y) - \alpha] + \varphi(\alpha) + B(x, y), \qquad \text{(XX.16,19)}$$

unde $A(x, y)$ şi $B(x, y)$ sînt două funcţii de x şi y, iar $\varphi(\alpha)$ o funcţie de α, care admite derivata $\varphi'(\alpha)$. Deoarece $f(x, y; \alpha)$ este o funcţie de frecvenţă, trebuie să avem

$$\int_{-\infty}^{+\infty} e^{\varphi'(\alpha)[A(x, y) - \alpha] + \varphi(\alpha) + B(x, y)}\, dy = 1.$$

Derivînd în raport cu α, rezultă

$$\int_{-\infty}^{+\infty} [A(x,\ y)-\alpha]\, f(x,\ y;\ \alpha)\, dy=0 \qquad \text{(XX.16,20)}$$

sau

$$M[A(x,\ y)]=\alpha, \qquad \text{(XX.16,21)}$$

variabila aleatorie fiind y, iar x un parametru.
De aici rezultă că

$$M[\bar{\alpha}(x_1,\ x_2,\ \ldots,\ x_n)]=\frac{\sum_{i=1}^{n} M[A(x_{i-1},\ x_i)]}{n}=\alpha. \qquad \text{(XX.16,22)}$$

Derivînd relația (XX.16,20) în raport cu α, obținem mai departe

$$\int_{-\infty}^{+\infty} [A(x,\ y)-\alpha]^2 f(x,\ y;\ \alpha)\, dy=\frac{1}{\varphi''(\alpha)}.$$

Pe de altă parte, deducem din (XX.16,18)

$$D^2[\bar{\alpha}(x_1,\ x_2,\ \ldots,\ x_n)]=\frac{\sum_{i=1}^{n} D^2[A(x_{i-1},\ x_i)]}{n^2}=\frac{1}{n\varphi''(\alpha)}, \qquad \text{(XX.16,23)}$$

deoarece

$$\iint\ldots\int [A(x_{i-1},\ x_i)-\alpha]\,[A(x_{j-1},\ x_j)-\alpha]\prod_{h=1}^{n} f(x_{h-1},\ x_h;\ \alpha)\, dx_h=0$$
$$(i\neq j).$$

Pe de altă parte

$$M\left[\frac{\partial \ln f(x_{i-1},\ x_i;\ \alpha)}{\partial\alpha}\right]^2=\varphi''^2(\alpha)\, M[A(x_{i-1},\ x_i)-\alpha]^2=\varphi''(\alpha),$$

$$\sum_{i=1}^{n} M\left[\frac{\partial \ln f(x_{i-1},\ x_i;\ \alpha)}{\partial\alpha}\right]^2=n\varphi''(\alpha). \qquad \text{(XX.16,24)}$$

Inlocuind valorile date de relațiile (XX.16,23) și (XX.16,24) în (XX.16,7), se vede că această inegalitate se transformă în egalitate. Prin urmare, condițiile (XX.16,19) sînt și suficiente.
Am demonstrat prin urmare următoarea

Teoremă. *Să considerăm selecția*

$$x_1, x_2, \ldots, x_n,$$

variabilele x_i $(i=1, 2, \ldots, n)$ *succedîndu-se după legea unui lanț Markov simplu și constant.*

Notăm prin $f(x, y; \alpha)\,dy$ *probabilitatea ca să avem* $y<x_p<y+dy$, *condiționată de* x_{p-1}, *pentru* $p=2, 3, \ldots, n$. *Notăm prin* $f(x_0, y; \alpha)\,dy$ *probabilitatea ca să avem* $y<x_1<y+dy$.

Admitem că

1) $f(x, y; \alpha)>0$.

2) $f(x, y; \alpha)$ *este continuă în raport cu* x, y, α *și admite derivată de ordinul întîi în raport cu* α.

Dacă $\bar{\alpha}(x_1, x_2, \ldots, x_n)$ *este o funcție de estimație absolut corectă a lui* α, *avem*

$$D^2[\bar{\alpha}(x_1, x_2, \ldots, x_n)] \geqslant \frac{1}{\sum_{i=1}^{n} M\left[\frac{\partial \ln f(x_{i-1}, x_i; \alpha)}{\partial \alpha}\right]^2}.$$

In relația de mai sus, avem egalitate dacă

$$\ln f(x, y; \alpha)=\varphi'(\alpha)\,[A(x, y)-\alpha]+\varphi(\alpha)+B(x, y),$$

unde $A(x, y)$, $B(x, y)$ *sînt două funcții de* x *și* y, *iar* $\varphi(\alpha)$ *este o funcție de* α, *avînd derivata* $\varphi'(\alpha)$.

Să considerăm ca exemplu lanțul Markov cu probabilitatea de trecere

$$f(x, y; \alpha)=\frac{1}{\sqrt{2\pi}\alpha} e^{-\frac{(y-\,)^2}{2\alpha^2}}\, dy.$$

Avem îndeplinită condiția

$$\frac{1}{\sqrt{2\pi}\alpha}\int_{-\infty}^{+\infty} e^{-\frac{(y-x)^2}{2\alpha^2}}\, dy=1,$$

$$\ln f(x, y; \alpha)=-\frac{(y-x)^2}{2\alpha^2}-\ln\alpha-\ln\sqrt{2\pi}.$$

Egalitatea (XX.16,19) nu este îndeplinită. Deci nu putem găsi o funcție de estimație eficientă pentru α.

Putem obține însă o funcție de estimație eficientă pentru α^2. Notînd $\alpha^2=\beta$, vom avea

$$\ln f(x,\ y;\ \alpha)=-\frac{(y-x)^2}{2\beta}-\frac{1}{2}\ln\beta-\ln\sqrt{2\pi}.$$

Putem lua

$$A(x,\ y)=(y-x)^2,\quad \varphi(\beta)=-\frac{1}{2}\ln\beta.$$

Condițiile din teorema precedentă sînt îndeplinite și prin urmare

$$\bar{\alpha}(x_1,\ x_2,\ \ldots,\ x_n)=\frac{1}{n}\sum_{i=1}^{n}(x_i-x_{i-1})^2$$

este o funcție de estimație eficientă pentru α^2.
Intr-adevăr, avem

$$\iint\ldots\int(\xi_i-\xi_{i-1})^2\prod_{i=1}^{n}f(\xi_{j-1},\ \xi_j;\ \alpha)\,d\xi_j=$$

$$=\iint\ldots\int\prod_{j=1}^{i-1}f(\xi_{j-1},\ \xi_j;\ \alpha)\,d\xi_j\int(\xi_i-\xi_{i-1})^2 f(\xi_{i-1},\ \xi_i;\ \alpha)\,d\xi_i=\alpha^2,$$

deoarece

$$\int(\xi_{i-1}-\xi_i)^2 f(\xi_{i-1},\ \xi_i;\ \alpha)\,d\xi_i=\frac{1}{\sqrt{2\pi}\alpha}\int(\xi_i-\xi_{i-1})^2 e^{-\frac{(\xi_i-\xi_{i-1})^2}{2\alpha^2}}\,d\xi_i=\alpha^2,$$

oricare ar fi valoarea lui ξ_{i-1}.
Rezultă imediat că

$$\frac{1}{n}\iint\ldots\int\sum_{i=1}^{n}(\xi_i-\xi_{i-1})^2\prod_{j=1}^{n}f(\xi_{j-1},\ \xi_j;\ \alpha)\,d\xi_j=\alpha^2.$$

Prin urmare

$$D^2(\bar{\alpha})=\frac{1}{n}M\left[\sum_{i=1}^{n}(x_i-x_{i-1})^2\right]=\alpha^2.$$

Pe de altă parte

$$\left[\frac{\partial\ln f(x_{i-1},\ x_i;\ \alpha)}{\partial(\alpha^2)}\right]^2=\frac{1}{2\alpha^8}\left[(x_i-x_{i-1})^4-2\alpha^2(x_i-x_{i-1})^2+\alpha^4\right]$$

și repetînd calculul de mai sus

$$\sum_{i=1}^{n} M\left[\frac{\partial \ln f(x_{i-1}, x_i; \alpha)}{\partial(\alpha^2)}\right]^2 = \frac{1}{\bar{\alpha}^2}.$$

Avem deci

$$D^2(\bar{\alpha}) = \frac{1}{\sum_{i=1}^{n} M\left[\frac{\partial \ln f(x_{i-1}, x_i; \alpha)}{\partial(\alpha^2)}\right]^2}$$

și deci $\bar{\alpha}$ este o funcție de estimație eficientă pentru α^2.

Extinzînd noțiunile din teoria estimației pentru variabilele independente și în cîmpul variabilelor dependente, putem atunci afirma că pentru α putem găsi o funcție de estimație suficientă.

§ 17. Cazul unui lanț Markov cu două stări, cu un singur parametru de estimat

Rezultatele obținute pentru lanțul Markov continuu sînt aplicabile și pentru lanțul Markov discontinuu. Să considerăm un lanț Markov simplu, constant, cu două stări, probabilitățile de trecere fiind elementele matricei stochastice

$$\left\| \begin{matrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{matrix} \right\|,$$

unde

$$p_{11}+p_{12}=1, \quad p_{21}+p_{22}=1.$$

Valorile luate de lanț sînt a_1 și a_2.

Avem

$$p_{ij} = P[x_n = a_j \mid x_{n-1} = a_i],$$

prin x_n $(n=1, 2, \ldots)$ notînd variabilele lanțului.

Să notăm

$$p_{12}=\lambda, \qquad p_{21}=\mu,$$
$$p_{11}=1-\lambda, \quad p_{22}=1-\mu$$

și să presupunem că, μ fiind cunoscut, avem să estimăm pe λ în baza unei selecții formate din n variabile consecutive ale lanțului

$$x_1, x_2, \ldots, x_n. \qquad \text{(XX.17,1)}$$

Să însemnăm prin x_0 o valoare inițială care precede pe x_1 și care poate lua una dintre valorile a_1, a_2. De asemenea, să însemnăm prin y_i variabila formată din perechile de variabile

$$(x_{i-1},\ x_i)\quad (i=1,\ 2,\ \ldots,\ n)$$

din șirul (XX.17,1).

Șirul $y_i\,(i=1,\ 2,\ \ldots,\ n)$ formează tot un lanț Markov simplu și constant, cu probabilitățile de trecere din matricea stochastică

$$\left\|\begin{matrix} p_{11} & p_{12} & 0 & 0 \\ 0 & 0 & p_{21} & p_{22} \\ p_{11} & p_{12} & 0 & 0 \\ 0 & 0 & p_{21} & p_{22} \end{matrix}\right\|.$$

Acest lanț ia valorile

$$b_1=(a_1,\ a_1),\quad b_2=(a_1,\ a_2),\quad b_3=(a_2,\ a_1),\quad b_4=(a_2,\ a_2),$$

iar elementul din matricea de mai sus, care se găsește în linia i și coloana j, reprezintă probabilitatea de a avea $y_s=b_j$, știind că $y_{s-1}=b_i$, oricare ar fi s.

Conform teoriei generale, pe care am expus-o în paragrafele precedente, pentru a găsi legea mixtă de probabilitate a numărului de cîte ori apar respectiv valorile b_1, b_2, b_3, b_4 în n probe consecutive, trebuie să considerăm determinantul

$$\Delta=\left\|\begin{matrix} p_{11}-1 & p_{12} & 0 & 0 \\ 0 & -1 & p_{21} & p_{22} \\ p_{11} & p_{12} & -1 & 0 \\ 0 & 0 & p_{21} & p_{22}-1 \end{matrix}\right\|.$$

Să notăm prin Q_j minorul elementului cuprins în linia j și coloana j din acest determinant și prin Q_{ij} minorul obținut din Δ prin suprimarea liniilor și coloanelor de rangul i și j.

Scanned with OKEN Scanner

Am văzut că funcția caracteristică a numărului de apariții ale valorilor b_1, b_2, b_3, b_4 este

$$\varphi(u)=e^{\frac{1}{2}\left(\sum_{j=1}^{4} P_j\xi_j^2+2\sum_{\substack{j,l=1\\(j\neq l)}}^{4} P_{jl}\xi_j\xi_l\right)}, \qquad \text{(XX.17,2)}$$

unde

$$P_j=\frac{Q_h}{\sum_{h=1}^{4} Q_h} \qquad (h=1,\ 2,\ 3,\ 4),$$

$$P_{hk}=-\frac{Q_{hk}}{\sum_{j=1}^{4} Q_j} \qquad (h, k=1,\ 2,\ 3,\ 4),$$

$$\xi_j=u_j-\sum_{h=1}^{4} P_j u_h \qquad (j=1,\ 2,\ 3,\ 4),$$

iar u_1, u_2, u_3, u_4 sînt variabilele funcției caracteristice.
Ținînd seamă de expresia lui Δ, obținem valorile

$$Q_1=-p_{11}p_{21},\quad Q_2=-p_{12}p_{21},\quad Q_3=-p_{12}p_{21},\quad Q_4=-p_{12}p_{22},$$
$$Q_{12}=p_{21},\quad Q_{13}=p_{21},\quad Q_{14}=1-p_{12}p_{21},\quad Q_{23}=p_{12}p_{21},$$
$$Q_{24}=p_{12},\quad Q_{34}=p_{12},$$

$$P_1=\frac{p_{11}p_{12}}{p_{12}+p_{21}},\quad P_2=\frac{p_{12}p_{21}}{p_{12}+p_{21}},\quad P_3=\frac{p_{12}p_{21}}{p_{12}+p_{21}},\quad P_4=\frac{p_{12}p_{22}}{p_{12}+p_{21}},$$

$$P_{12}=\frac{p_{21}}{p_{12}+p_{21}},\quad P_{13}=\frac{p_{21}}{p_{12}+p_{21}},\quad P_{14}=\frac{1-p_{12}p_{21}}{p_{12}+p_{21}},$$

$$P_{23}=\frac{p_{12}p_{21}}{p_{12}+p_{21}},\quad P_{24}=\frac{p_{12}}{p_{12}+p_{21}},\quad P_{34}=\frac{p_{12}}{p_{12}+p_{21}}.$$

Să însemnăm prin α_i $(i=1,\ 2,\ 3,\ 4)$ numărul respectiv de apariții ale valorilor b_i $(i=1,\ 2,\ 3,\ 4)$ în selecția y_i $(i=1,\ 2,\ \ldots,\ n)$.

Din expresia funcției caracteristice (XX.17,2) rezultă imediat

$$D^2(\alpha_1) \simeq \frac{\lambda\mu(1-\lambda)(2-\lambda+\mu+\lambda\mu+\mu^2)}{(\lambda+\mu)^3} n,$$

$$D^2(\alpha_2) \simeq \frac{\lambda\mu(\lambda^2+\mu^2-\lambda\mu^2-\lambda^2\mu)}{(\lambda+\mu)^3} n, \qquad \text{(XX.17,3)}$$

$$D^2(\alpha_3) \simeq \frac{\lambda\mu(\lambda^2+\mu^2-\lambda^2\mu)}{(\lambda+\mu)^3} n$$

$$D^2(\alpha_4) \simeq \frac{\lambda\mu(1-\mu)(2+\lambda-\mu+\lambda^2+\mu^2)}{(\lambda+\mu)^3} n,$$

$$M[(\alpha_1-P_1)(\alpha_3-P_3)] \simeq \frac{\lambda\mu[-\lambda+\mu+\lambda^2+\mu^2-\lambda^2\mu-\lambda\mu^2]}{(\lambda+\mu)^3} n.$$

La fel

$$\lim_{n\to\infty} M\left(\frac{\alpha_1}{n}\right) = \frac{(1-\lambda)\mu}{\lambda+\mu},$$

$$\lim_{n\to\infty} M\left(\frac{\alpha_2}{n}\right) = \frac{\lambda\mu}{\lambda+\mu}, \qquad \text{(XX.17,4)}$$

$$\lim_{n\to\infty} M\left(\frac{\alpha_3}{n}\right) = \frac{\lambda\mu}{\lambda+\mu},$$

$$\lim_{n\to\infty} M\left(\frac{\alpha_4}{n}\right) = \frac{\lambda(1-\mu)}{\lambda+\mu}.$$

Pentru estimarea lui p_{11} vom alege funcția de estimație

$$\frac{\alpha_1}{\alpha_1+\alpha_3}.$$

Din formulele precedente rezultă

$$\lim_{n\to\infty} M\left(\frac{\alpha_1+\alpha_3}{n}\right) = \frac{\mu}{\lambda+\mu}, \qquad \text{(XX.17,5)}$$

$$D^2(\alpha_1+\alpha_3) \simeq \frac{\lambda\mu(2-\lambda-\mu)}{(\lambda+\mu)^3} n,$$

$$M[\alpha_1-M(\alpha_1)][\alpha_1+\alpha_3-M(\alpha_1+\alpha_3)] \simeq \frac{2\lambda\mu(1-\lambda)}{(\lambda+\mu)^3} n.$$

Scanned with OKEN Scanner

Raportul $\frac{\alpha_1}{\alpha_1+\alpha_3}$ rămîne mărginit, oricare ar fi valorile luate de α_1 și α_3, fiind întotdeauna mai mic decît unu.

Atunci, după o teoremă cunoscută din calculul probabilităților, care ne dă valoarea medie și dispersia unui raport de două variabile în funcție de primele două momente ale acestor variabile, avem

$$M\left(\frac{\alpha_1}{\alpha_1+\alpha_2}\right)=\frac{M(\alpha_1)}{M(\alpha_1+\alpha_3)}+O\left(\frac{1}{n}\right),$$

$$D^2\left(\frac{\alpha_1}{\alpha_1+\alpha_3}\right)=\frac{M^2(\alpha_1)}{nM^2(\alpha_1+\alpha_3)}\left[\frac{D^2(\alpha_1)}{M^2(\alpha_1)}-\frac{2M(\alpha_1-P_1)(\alpha_1+\alpha_3-P_1-P_3)}{M(\alpha_1)M(\alpha_1+\alpha_3)}+\right.$$

$$\left.+\frac{D^2(\alpha_1+\alpha_2)}{M^2(\alpha_1+\alpha_3)}\right]+O\left(n^{-\frac{3}{2}}\right).$$

Ținînd seama de formulele (XX.17,4) și (XX.17,5) obținem astfel

$$D^2\left(\frac{\alpha}{\alpha_1+\alpha_3}\right)=\frac{\lambda(1-\lambda)(\lambda+\mu)}{n\mu}+O\left(n^{-\frac{3}{2}}\right), \qquad \text{(XX.17,6)}$$

$$M\left(\frac{\alpha_1}{\alpha_1+\alpha_3}\right)=1-\lambda+O\left(n^{-\frac{3}{2}}\right)=p_{11}+O\left(\frac{1}{n}\right).$$

Aceste relații ne arată că $\frac{\alpha_1}{\alpha_1+\alpha_3}$ este o estimație corectă a lui p_{11}.

Se poate arăta că această funcție de estimație a lui p_{11} este asimptotic eficientă.

Intr-adevăr, avem

$$\frac{\partial \ln p_{11}}{\partial\lambda}=-\frac{1}{1-\lambda},\quad \frac{\partial \ln p_{12}}{\partial\lambda}=\frac{1}{\lambda},$$

$$\frac{\partial \ln p_{21}}{\partial\lambda}=0,\quad \frac{\partial \ln p_{22}}{\partial\lambda}=0,$$

$$\sum_{i=1}^{n} M\left[\frac{\partial \ln f(x_{i-1},\, x_i;\, \lambda)}{\partial\lambda}\right]^2=\frac{1}{\lambda(1-\lambda)}\sum_{j=1}^{n}P_i^{(j)},$$

unde prin $P_i^{(j)}$ notăm probabilitatea ca să avem $x_j=a_i$.

Pe de altă parte, am văzut din teoria lanțurilor lui Markov că

$$\lim_{n\to\infty}\sum_{j=1}^{n} P_i^{(j)} = \frac{\mu}{\lambda+\mu}.$$

Obținem astfel

$$\sum M\left[\frac{\partial \ln f(x_{i-1}, x_i; \lambda)}{\partial\lambda}\right]^2 \simeq \frac{n\mu}{\lambda(1-\lambda)(\lambda+\mu)}.$$

§ 18. Cazul unui lanț Markov cu mai mulți parametri

Am văzut că, dacă selecția este formată dintr-o succesiune de variabile înlănțuite Markov în mod simplu, probabilitățile de trecere depinzînd de un singur parametru, pentru estimarea acestui parametru dispersiile funcțiilor de estimație admit o margine inferioară, diferită de zero.

In cazul cînd această margine este atinsă, funcția de estimație care se bucură de această proprietate poartă numele de *funcție de estimație eficientă*, dacă dispersia ei tinde către marginea inferioară, cînd mărimea selecției crește necontenit. Aceste proprietăți reprezintă o extensiune în domeniul lanțurilor Markov a teoremelor fundamentale din teoria estimației pentru variabilele independente.

Proprietățile privitoare la minimul dispersiei pentru cazul cînd în distribuția teoretică există mai mulți parametri se păstrează și pentru lanțurile Markov. Demonstrația este analogă cu aceea dată pentru cazul unui singur parametru. De aceea nu insistăm asupra ei, limitîndu-ne numai la considerarea unui exemplu, care prezintă interes pentru el însuși.

Să considerăm un lanț Markov cu m stări a_i $(i=1, 2, \ldots, m)$. Lanțul este simplu și constant. Fie p_{ij} probabilitățile de trecere corespunzătoare. Insemnînd prin

$$x_1, x_2, \ldots, x_n \qquad \text{(XX.18.1)}$$

o succesiune de n variabile ale lanțului, care formează o selecție de mărime n, avem deci

$$p_{ij} = P(x_k = a_j \mid x_{k-1} = a_i) \quad (i, j = 1, 2, \ldots, m)$$

oricare ar fi k $(k=1, 2, \ldots, n)$.

Probabilitățile de trecere îndeplinesc condițiile

$$p_{ij} \geqq 0, \quad \sum_{j=1}^{m} p_{ij} = 1. \tag{XX.18,2}$$

Ne propunem să estimăm probabilitățile p_{ij} $(i, j = 1, 2, \ldots, m)$ în baza rezultatelor date de selecția (XX.18,1). Ținînd seamă de (XX.18,2) înseamnă că avem $m(m-1)$ parametri de estimat.

Să notăm prin n_j numărul de apariții al stării a_j în selecția considerată și prin n_{ij} numărul de apariții al stării a_j precedată de starea a_i.

Avem

$$\sum_{j=1}^{m} n_j = n, \quad \sum_{i=1}^{m} n_{ij} = n_j.$$

V. Romanovski (1949) și N. Smirnov (1955) au ales pentru estimarea probabilității p_{ij} funcția de estimație $\frac{n_{ij}}{n_j}$. Ei s-au ocupat de verificarea ipotezei că aceste probabilități au anumite valori, întrebuințînd pentru aceasta criteriul χ^2 al lui Pearson.

Vom trata, în cele ce urmează, problema dintr-un alt punct de vedere și anume vom arăta că funcțiile de estimație considerate sînt *asimptotic eficiente*.

Ne propunem, în primul rînd, să calculăm valoarea medie și dispersia variabilei n_{ij}.

Pentru aceasta, reamintim următoarele date rezultate din expunerea pe care am făcut-o asupra variabilelor vectoriale înlănțuite în sensul lui Markov.

Considerăm determinantul

$$\Delta = |p_{ik} - \delta_{ik}| \qquad (i, k = 1, 2, \ldots, m),$$

unde $\delta_{ik} = 1$ pentru $i = k$ și $\delta_{ik} = 0$, pentru $i \neq k$. Insemnăm prin Q_h minorul elementului $p_{hh} - 1$ din Δ și prin Q_{ij} minorul aceluiași determinant, obținut prin suprimarea liniilor și coloanelor de rang i și j.

Să mai însemnăm

$$P_h = \frac{Q_h}{\sum_{j=1}^{m} Q_j}, \quad P_{hk} = -\frac{Q_{hk}}{\sum_{j=1}^{m} Q_j}.$$

Dacă

$$\xi_h = u_h - \sum_{j=1}^{m} P_j u_j,$$

am văzut că funcția caracteristică corespunzătoare numărului de treceri prin fiecare stare, raportat la $\sqrt{n}$, tinde pentru $n \to \infty$, către

$$e^{\frac{1}{2}\left(\sum_{j=1}^{m} P_j \xi_j^2 + 2 \sum_{j,l=1}^{m} P_{jl} \xi_j \xi_l\right)}. \qquad \text{(XX.18,3)}$$

Considerînd coeficientul lui u_j și u_j^2 din funcția caracteristică limită rezultă că

$$M\left(\frac{n_j}{n}\right) = P_j + \varphi_j(n), \qquad \text{(XX.18,4)}$$

$$D^2\left(\frac{n_j}{n}\right) = -P_j + P_j^2 + 2P_j \sum_{h=1}^{m} P_{jh} - 2P_j^2 \sum_{h,k=1}^{m} P_{hk} + \psi_j(n), \qquad \text{(XX.18,5)}$$

unde $\varphi_j(n)$ și $\psi_j(n)$ sînt două funcții de n, care tind către zero, cînd $n \to \infty$.

In mod analog, însemnînd prin $c(uv)$ covarianța variabilelor u și v, rezultă

$$c\left(\frac{n_i}{n}\frac{n_j}{n}\right) = -P_{ij} + P_i P_j + P_i \sum_{h=1}^{m} P_{ih} +$$

$$+ P_j \sum_{h=1}^{m} P_{jh} - 2P_i P_j \sum_{h,k=1}^{m} P_{hk} + \chi_{ij}(n), \qquad \text{(XX.18,6)}$$

unde $\chi_{ij}(n)$ este o funcție de n care tinde către zero, cînd n tinde către infinit.

Putem calcula valoarea medie și dispersia lui n_{ij} în același mod ca în paragraful precedent, adică făcînd o diferențiere între stările a_j, după stările care o preced.

Putem simplifica însă calculele, folosind o altă metodă.

Am arătat la cap. IV. § 4 că

$$M(n_{ij}) = M[M(n_{ij} \mid n_j)] \qquad \text{(A)}$$

$$D^2(n_{ij}) = D^2[M(n_{ij} \mid n_j)] + M[D^2(n_{ij} \mid n_j)]. \qquad \text{(B)}$$

Rămîne să determinăm variabilele $M(n_{ij} \mid n_i)$ și $D^2(n_{ij} \mid n_i)$.

Să notăm prin $P(n_j)$ probabilitatea ca în selecția (XX.18,1), starea a_j să se prezinte de n_j ori. Este evident că avem $0 \leqslant n_j \leqslant n$ și că

$$\sum_{n_j=0}^{n} P(n_j) = 1,$$

$$\sum_{n_j=0}^{n} P(n_j)\, n_j \simeq nP_j, \qquad \text{(XX.18,7)}$$

$$\sum_{n_j=0}^{n} P(n_j)\,(n_j - n\,P_j)^2 \simeq$$

$$\simeq n\left[-P_j + P_j^2 + P_j \sum_{h=1}^{m} P_{jh} - 2P_j^2 \sum_{h,\,k=1}^{m} P_{hk}\right]. \qquad \text{(XX.18,8)}$$

Ultimele două formule rezultă din (XX.18,4) și (XX.18,5) și ne dau valoarea asimptotică a sumelor considerate în membrul întîi.

Fie

$$x_{\xi_1},\ x_{\xi_2},\ \ldots,\ x_{\xi_{n_i}}$$

variabilele din selecția (XX.18,1) care au trecut prin stările

$$a_i\,(\xi_1 < \xi_2 < \ldots < \xi_{n_i}).$$

Numerele $\xi_1,\ \xi_2,\ \ldots,\ \xi_{n_i}$ sînt aleatorii.

Fie Y_k grupul de variabile aleatorii $x_{\xi_k+1},\ x_{\xi_k+2},\ \ldots,\ x_{\xi_{k+1}}$, unde $k = 1,\ 2,\ \ldots,\ n_i$. Variabilele Y_k sînt independente între ele.

În selecția (XX.18,1) starea a_j va fi precedată de starea a_i numai dacă într-o variabilă Y_k avem $x_{\xi_k+1} = a_j$, ceea ce se întîmplă, dat fiind faptul că $x_{\xi_k} = a_i$, cu probabilitatea p_{ij}. Prin urmare, dacă știm că starea a_i s-a prezentat de n_i ori, n_{ij} reprezintă numărul de ori de care s-a verificat un eveniment de probabilitate constantă P_{ij} în n_i probe independente. Avem, deci, aplicînd formulele cunoscute din distribuția binomială

$$M(n_{ij} \mid n_i) = n_i p_{ij}, \qquad \text{(XX.18,9)}$$

$$D^2(n_{ij} \mid n_i) = n_i p_{ij}(1 - p_{ij}). \qquad \text{(XX.18,10)}$$

Valorile (XX.18,9) și (XX.18,10) sînt luate respectiv cu probabilitățile $P(n_i)$.
Avem atunci

$$M[M(n_{ij}\mid n_i)]=\sum_{n_i=0}^{n} P(n_i)\, n_i p_{ij},$$

$$M[D^2(n_{ij}\mid n_i)]=\sum_{n_i=0}^{n} P(n_i)\, n_i p_{ij}\,(1-p_{ij}).$$

Ținînd seamă de (XX.18,7), relațiile de mai sus devin

$$M[M(n_{ij}\mid n_i)]=np_{ij}P_i, \qquad \text{(XX.18,11)}$$

$$M[D^2(n_{ij}\mid n_i)]=np_{ij}(1-p_{ij})P_i. \qquad \text{(XX.18,12)}$$

Ținînd seamă de formula (A) rezultă

$$M(n_{ij})\approx nP_i p_{ij}. \qquad \text{(XX.18,13)}$$

Pentru a aplica formula (B), mai rămîne să calculăm expresia $D^2[M(n_{ij}\mid n_i)]$. Am văzut însă că $M(n_{ij}\mid n_i)$ este o variabilă aleatorie care ia valorile $n_i p_{ij}$ $(n_i=0,\ 1,\ \ldots,\ n)$ cu probabilitățile respective $P(n_i)$. Valoarea medie a acestei variabile am văzut că este $nP_i p_{ij}$. Prin urmare

$$M[D^2(n_{ij}\mid n_i)]=\sum_{n_i=0}^{n} P(n_i)\,[n_i p_{ij}-nP_i p_{ij}]^2=$$

$$=p_{ij}^2\sum_{n_i=0}^{n} P(n_i)\,[n_i-nP_i]^2$$

sau, ținînd seamă de relațiile (XX.18,8)

$$M[D^2(n_{ij}\mid n_i)]=$$

$$=np_{ij}^2\left[-P_i+P_i^2+2P_i\sum_{h=1}^{m}P_{ih}-2P_i^2\sum_{h,\,k=1}^{m}P_{hk}\right]. \qquad \text{(XX.18,14)}$$

Inlocuind în formula (B) valorile date prin relațiile (XX.18,12) și (XX.18,14), obținem

$$D^2(n_{ij}) \approx$$

$$\approx n\left[p_{ij}P_i - 2p_{ij}^2P_i + P_{ij}^2P_i^2 + 2p_{ij}^2P_i \sum_{h=1}^{m} P_{ih} - 2p_{ij}^2 \sum_{h,\,k=1}^{m} P_{hk}\right]. \quad \text{(XX.18,15)}$$

In același mod, putem calcula covarianța dintre n_{ih} și n_{ik}, plecînd de la formula stabilită la cap. IV, § 3

$$C(n_{ih}n_{ik}) = C[M(n_{ih} \mid n_i)\, M(n_{ik} \mid n_i)] + M[C(n_{ih}n_{ik} \mid n_i)]. \quad \text{(C)}$$

Starea a_h va fi precedată, în selecția considerată, de a_i numai dacă într-o variabilă Y_k vom avea $x_{\xi_k+1} = a_h$, ceea ce se întîmplă cu probabilitatea p_{ih}, în condiții analoge.

Pentru fiecare variabilă Y_k avem probabilitatea p_{ih} ca să se ivească starea a_h înainte de a_i, probabilitatea p_{ik} ca să se producă a_k înainte de a_i și probabilitatea $1-p_{ih}-p_{ik}$ pentru ca să nu se producă nici unul dintre aceste evenimente. Prin urmare, n_{ih} și n_{ik} reprezintă numărul respectiv de cîte ori se ivesc în n_i probe, la distribuția multinomială cu trei stări, avînd probabilitățile respective p_{ih}, p_{ik}, $1-p_{ih}-p_{ik}$, primele două stări, dacă știm bineînțeles că starea a_i s-a produs de n_i ori.

Avem în acest caz

$$M(n_{ih} \mid n_i) = n_i p_{ih}, \quad M(n_{ik} \mid n_i) = np_{ik}, \quad \text{(XX.18,16)}$$

$$D^2(n_{ih} \mid n_i) = n_i p_{ih}(1-p_{ih}), \quad D^2(n_{ik} \mid n_i) = n_i p_{ik}(1-p_{ik}), \quad \text{(XX.18,17)}$$

$$C(n_{ih}n_{ik} \mid n_i) = -n_i p_{ih} p_{ik}. \quad \text{(XX.18,18)}$$

Toate aceste valori sînt luate cu probabilitatea $P(n_i)$.

Rezultă că

$$M[C(n_{ih}n_{ik} \mid n_i)] = -\sum_{n_i=0}^{n} P(n_i)\, n_i p_{ih} p_{ik} = -nP_i p_{ih} p_{ik}, \quad \text{(XX.18,19)}$$

Pe de altă parte

$$C\left[M\left(n_{ih}\mid n_i\right),\ M\left(n_{ik}\mid n_i\right)\right]=\sum_{n_i=0}^{n}P\left(n_i\right)\left[n_ip_{ih}-nP_ip_{ih}\right]\left[n_ip_{ik}-\right.$$

$$\left.-nP_ip_{ik}\right]=p_{ih}p_{ik}\sum_{n_i=0}^{n}P\left(n_i\right)\left(n_i-nP_i\right)^2\approx$$

$$\approx np_{ih}p_{ik}\left[-P_i+P_i^2+2P_i\sum_{h=1}^{m}P_{ih}-2P_i^2\sum_{h,k=1}^{m}P_{hk}\right]. \qquad \text{(XX.18,20)}$$

Introducînd formulele (XX.18,19) şi (XX.18,20) în (C) rezultă

$$C\left(n_{ih}n_{ik}\right)\approx$$

$$\approx n\left[-2p_{ih}p_{ik}P_i+p_{ih}p_{ik}P_i^2+2p_{ih}p_{ik}P_i\sum_{h=1}^{m}P_{ih}-2p_{ih}p_{ik}P_i^2\sum_{h,k=1}^{m}P_{hk}\right]. \qquad \text{(XX.18,21)}$$

Dacă ţinem seamă de faptul că

$$n_i=\sum_{j=1}^{n}n_{ij},$$

şi de relaţiile (XX.18,15) şi (XX.18,21) rezultă imediat

$$C\left(n_in_{ij}\right)\approx$$

$$\approx n\left[-p_{ij}P_i+p_{ij}P_i^2+2p_{ii}P_i\sum_{h=1}^{m}P_{ih}-2p_{ii}P_i^2\sum_{h,k=1}^{m}P_{hk}\right]. \qquad \text{(XX.18,22)}$$

Sîntem acum în măsură să calculăm valoarea medie şi dispersia raportului $\frac{n_{ij}}{n_i}$.

In adevăr, sîntem în condiţiile teoremei pe care am demonstrat-o la § 16.

Prin urmare, putem scrie

$$M\left(\frac{n_{ij}}{n_i}\right)\approx\frac{M\left(n_{ij}\right)}{M\left(n_i\right)}.$$

Scanned with OKEN Scanner

Ținînd seamă de (XX.18,4) și (XX.18,11), rezultă

$$M\left(\frac{n_{ij}}{n_i}\right)\approx p_{ij}. \qquad \text{(XX.18,23)}$$

Deoarece

$$M(n_{ij})\approx nP_i p_{ij}, \quad M(n_i)\approx np_i,$$

avem, în baza teoremei menționate,

$$D^2\left(\frac{n_{ij}}{n_i}\right)\approx\frac{p_{ij}^2}{n^2}\left[\frac{D^2(n_{ij})}{P_i^2 p_{ij}^2}-\frac{2C(n_i n_{ij})}{P_i^2 p_{ij}}+\frac{D^2(n_i)}{P_i^2}\right].$$

Ținînd seamă de relațiile (XX.18,5), (XX.18,15), (XX.18,22), egalitatea de mai sus ne dă

$$D^2\left(\frac{n_{ij}}{n_i}\right)\approx\frac{p_{ij}(1-p_{ij})}{nP_i}. \qquad \text{(XX.18,24)}$$

Din relațiile (XX.18,23) și (XX.18,24) reiese că raportul $\frac{n_{ij}}{n_i}$ este o funcție de estimație corectă pentru p_{ij}.

Să calculăm acum în funcție de elementele lanțului și covarianța variabilelor

$$\frac{n_{ih}}{n_i}, \quad \frac{n_{ik}}{n_i}.$$

Din relația

$$D^2(n_{ih}+n_{ik})=D^2(n_{ik})+2C(n_{ih}n_{ik})+D^2(n_{ik}),$$

deducem ținînd seamă de (XX.18,15) și (XX.18,21)

$$D^2(n_{ih}+n_{ik})\approx$$

$$\approx n[(p_{ih}+p_{ik})P_i-2(p_{ih}+p_{ik})^2P_i+(p_{ih}+p_{ik})^2P_i^2+$$

$$+2(p_{ih}+p_{ik})^2\sum P_{ih}-2(p_{ih}+p_{ik})^2\sum P_{hk}]. \qquad \text{(XX.18,25)}$$

Se vede că pentru a obține dispersia sumei $n_{ih}+n_{ik}$ este suficient în formula (XX.18,15) să înlocuim pe p_{ij} prin $p_{ih}+p_{ik}$, ceea ce rezultă direct și fără calcul.

Repetînd același raționament, obținem

$$D^2\left(\frac{n_{ih}+n_{ik}}{n_i}\right)\approx\frac{(p_{ih}+p_{ik})(1-p_{ih}-p_{ik})}{nP_i}.$$

Acum, din formula

$$D^2\left(\frac{n_{ih}+n_{ik}}{n_i}\right)=D^2\left(\frac{n_{ih}}{n_i}\right)+2C\left(\frac{n_{ih}}{n_i}\frac{n_{ik}}{n_i}\right)+D^2\left(\frac{n_{ik}}{n_i}\right),$$

ținînd seamă de egalitatea precedentă și de (XX.18,15), rezultă imediat

$$C\left(\frac{n_{ih}}{n_i}\frac{n_{ik}}{n_i}\right)=-\frac{p_{ih}p_{ik}}{nP_i}. \qquad \text{(XX.18,26)}$$

După cum se vede, formulele (XX.18,24) și (XX.18,25) sînt simple. Dacă notăm $n_i=nP_i$, ele coincid cu dispersiile și covarianțele apariților diferitelor stări în distribuția multinomială. Acest rezultat se aplică prin independența variabilelor Y_k mai sus considerate. Demonstrarea formulelor (XX.18,24) și (XX.18,26) se poate face de aceea și în alt mod, tolosind această observație.

Ne-am propus să estimăm probabilitățile de trecere p_{ij} pe baza selecției (XX.18,1) cu scopul de a vedea cum se aplică teorema minimului dispersiei funcțiilor de estimație în cazul variabilelor dependente, cînd sînt mai mulți parametri. Pentru simplificarea calculelor, vom presupune că parametrii necunoscuți sînt

$$p_{i1},\ p_{i2},\ldots,p_{im},$$

în număr de $(m-1)$ din cauza relației

$$\sum_{h=1}^{m}p_{ih}=1.$$

Vom nota

$$p_{ih}=\alpha_h,\quad h=1,\ 2,\ldots,m-1.$$

Avem

$$\sum_{j=1}^{m}\left(\frac{\partial\ln P_{ij}}{\partial\alpha_h}\right)^2 p_{ij}=\frac{p_{ih}+p_{im}}{p_{ih}p_{im}},$$

$$\sum_{j=1}^{m}\frac{\partial\ln p_{ij}}{\partial\alpha_h}\frac{\partial\ln p_{ij}}{\partial\alpha_k}=\frac{1}{p_{im}}.$$

Rezultă atunci, conform teoriei generale pe care am expus-o la § 17, că

$$M\left(\frac{\partial \ln P}{\partial \alpha_h}\right)^2 = \frac{p_{ih}+p_{im}}{p_{ih}\,p_{im}} \sum_{k=1}^{n} P_i^{(k)} \simeq nP_i \frac{p_{ih}+p_{im}}{p_{ih}\,p_{im}}, \qquad \text{(XX.18,27)}$$

$$M\left(\frac{\partial \ln P}{\partial \alpha_h}\frac{\partial \ln P}{\partial \alpha_k}\right) = \frac{1}{p_{im}} \sum_{k=1}^{n} P_i^{(k)} \approx nP_i \frac{1}{p_{im}}. \qquad \text{(XX.18,28)}$$

Prin urmare, ecuația de concentrație minimă a dispersiei este

$$\sum_{h=1}^{m-1} \frac{p_{ih}+p_{im}}{p_{ih}\,p_{im}} (x_h - p_{ih})^2 + 2 \sum_{\substack{h,\,k=1 \\ (h \neq k)}}^{m-1} (x_h - p_{ih})(x_k - p_{ik}) = \frac{(m+1)\,p_{im}}{nP_i}. \qquad \text{(XX.18,29)}$$

Pe de altă parte, din formulele (XX.18,24) și (XX.18,26) se vede că avem de considerat matricea

$$A = \left\| \begin{array}{cccc} \frac{p_{i1}(1-p_{i1})}{nP_i} & -\frac{p_{i1}\,p_{i2}}{nP_i} & \cdots & -\frac{p_{i1}\,p_{i,m-1}}{nP_i} \\ -\frac{p_{i2}\,p_{i1}}{nP_i} & \frac{p_{i2}(1-p_{i2})}{nP_i} & \cdots & -\frac{p_{i2}\,p_{i,m-1}}{nP_i} \\ \cdots & \cdots & \cdots & \cdots \\ -\frac{p_{i,m-1}\,p_{i1}}{nP_i} & -\frac{p_{i,m-1}\,p_{i2}}{nP_i} & \cdots & \frac{p_{i,m-1}(1-p_{i,m-1})}{nP_i} \end{array} \right\|.$$

Un calcul elementar ne arată că

$$A^{-1} = \left\| \begin{array}{cccc} \frac{nP_i(p_{i1}+p_{im})}{p_{i1}\,p_{im}} & \frac{nP_i}{p_{im}} & \cdots & \frac{nP_i}{p_{im}} \\ \frac{nP_i}{p_{im}} & \frac{nP_i(p_{i2}+p_{im})}{p_{i2}\,p_{im}} & \cdots & \frac{nP_i}{p_{im}} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{nP_i}{p_{im}} & \frac{nP_i}{p_{im}} & \cdots & \frac{nP_i(p_{i,m-1}+p_{im})}{p_{i,m-1}\,p_{im}} \end{array} \right\|.$$

Elipsoidul de concentrație al funcțiilor de estimație n_{ih} $(h=1, 2,..., m-1)$ coincide, din punct de vedere asimptotic, cu elipsoidul de concentrație minimă al dispersiei.

Funcțiile de estimație considerate sînt prin urmare *asimptotic eficiente.*

§ 19. Metoda verosimilității maxime

Să considerăm din nou selecția (XX.18,1)

$$x_1, x_2, \ldots, x_n$$

și să însemnăm prin P probabilitatea obținerii valorilor (XX.18,1). Păstrînd notațiile din paragraful precedent, putem scrie

$$P=\Pi p_{ik}, \qquad \text{(XX.19,1)}$$

produsul aplicîndu-se la toate valorile n_{ik} întîlnite în selecția (XX.18,1). Admitem că parametrii necunoscuți sînt

$$p_{i1}=\alpha_1, \; p_{i2}=\alpha_2, \ldots, p_{i,\,m-1}=\alpha_{m-1}.$$

Din rezolvarea ecuației verosimilității maxime

$$\frac{\partial \ln P}{\partial \alpha_j}=0 \qquad (j=1, 2, \ldots, m-1)$$

rezultă

$$\alpha_j=\frac{n_{ij}}{n_j} \qquad (j=1, 2, \ldots, m-1). \qquad \text{(XX.19,2)}$$

Ajungem, prin urmare, la aceleași funcții de estimație ca în paragraful precedent. Am văzut la § 5, cap. XVII că, în cazul variabilelor independente, soluția ecuației verosimilității maxime, în condiții destul de generale, este o funcție de estimație eficient asimptotică.

Mai mult decît atît, această funcție de estimație este și normală. Se păstrează acest rezultat și pentru selecțiile dependente în sensul lui Markov? Rezultatele stabilite în paragraful precedent par să confirme acest lucru, deoarece am demonstrat direct că funcțiile de estimație (XX.19,2) sînt asimptotic eficiente.

Vom considera, pentru simplificare, cazul cînd distribuția teoretică dispune de un singur parametru.

Scanned with OKEN Scanner

Avem în acest caz

$$P(x_1, x_2, \ldots, x_n; \alpha) =$$

$$= f(x_0, x_1; \alpha) f(x_1, x_2; \alpha) \ldots f(x_{n-1}, x_n; \alpha), \quad \text{(XX.19,3)}$$

unde x_0 este o valoare inițială.

Rezultă

$$\frac{\partial \ln P(x_1, x_2, \ldots, x_n; \alpha)}{\partial \alpha} = \sum_{i=1}^{n} \frac{\partial \ln f(x_{i-1}, x_i; \alpha)}{\partial \alpha}. \quad \text{(XX.19,4)}$$

Ne propunem să arătăm că, pentru $n \to \infty$, suma din (XX.19,4), convenabil normată, tinde către distribuția normală.

Avem, după cum am arătat la § 15,

$$M\left[\frac{\partial \ln f(x_{i-1}, x_i; \alpha)}{\partial \alpha}\right] = 0, \quad \text{(XX.19,5)}$$

$$M\left[\sum_{i=1}^{n} \frac{\partial \ln f(x_{i-1}, x_i; \alpha)}{\partial \alpha}\right]^2 = \sum_{i=1}^{n} M\left[\frac{\partial \ln f(x_{i-1}, x_i; \alpha)}{\partial \alpha}\right]^2. \quad \text{(XX.19,6)}$$

Variabilele $\frac{\partial \ln f(x_{i-1}, x_i; \alpha)}{\partial \alpha}$ reprezintă un lanț Markov, simplu și constant.

Într-adevăr, dacă variabila $\frac{\partial \ln f(x_{i-1}, x_i; \alpha)}{\partial \alpha}$ a luat valoarea $\frac{\partial \ln f(x, y; \alpha)}{\partial \alpha}$, variabila următoare poate lua o valoare cuprinsă între $\frac{\partial \ln f(y, z; \alpha)}{\partial z}$ și $\frac{\partial \ln f(y, z+dz; \alpha)}{\partial z}$ cu o probabilitate egală cu $f(y, z; \alpha)\, dz$.

Admitem că funcția $|f(x, y; \alpha)|$ este mărginită în tot planul și că valorile luate de lanțul x_i $(i = 1, 2, \ldots, n)$ din selecția (XX.18,1) formează o singură mulțime finală. Rezultă atunci că lanțul mai sus definit admite de asemenea o singură mulțime finală.

Să admitem că $f(x, y; \alpha) > 0$, oricare ar fi x, y, α, și că este mărginită în tot planul (x, y), pentru oricare $\alpha \in \Delta$. Să mai facem în plus ipoteza că există un număr egal cu h, astfel ca

$$\int_{-\infty}^{+\infty} \left[\frac{\partial \ln f(x, y; \alpha)}{\partial \alpha}\right]^2 f(x, y; \alpha)\, dy > h^2 > 0. \quad \text{(XX.19,7)}$$

Din această condiție, rezultă imediat

$$\sum_{i=1}^{n} M\left[\frac{\partial \ln f(x_{i-1}, x_i; \alpha)}{\partial \alpha}\right]^2 > nh^2.$$

Sîntem în condițiile în care am demonstrat la § 15 că o sumă de variabile înlănțuite în sens Markov urmează la limită legea normală. Prin urmare, variabila

$$\xi = \frac{\sum_{i=1}^{n} \frac{\partial \ln f(x_{i-1}, x_i; \alpha)}{\partial \alpha}}{\sqrt{\sum_{i=1}^{n} M\left[\frac{\partial \ln f(x_{i-1}, x_i; \alpha)}{\partial \alpha}\right]^2}} \qquad \text{(XX.19,8)}$$

urmează, pentru $n \to \infty$, o distribuție normală. Putem extinde acum demonstrația dată de D. Dugué și care se găsește expusă în tratatul lui H. Cramér [1] și în cazul lanțului Markov și să arătăm că *ecuația verosimilității maxime*

$$\frac{\partial \ln P(x_1, x_2, \ldots, x_n; \alpha)}{\partial \alpha} = 0, \qquad \text{(XX.19,9)}$$

în condițiile pe care le vom arăta, *admite o soluție care reprezintă o funcție de estimație asimptotic normală și eficientă.*

Condițiile sînt următoarele

I. *Admitem existența primelor trei derivate ale funcției* $f(x, y; \alpha)$ *în raport cu* α, *oricare ar fi* x, y *și* α *cuprins într-un interval dat nedegenerat.*

II. *Pentru oricare* $\alpha \in \Delta$, *există funcțiile* $F_i(x, y)$ $(i = 1, 2, 3)$ *integrabile de la* $-\infty$ *la* $+\infty$ *în raport cu* y, *astfel ca*

$$\left|\frac{\partial f(x, y; \alpha)}{\partial \alpha}\right| < F_1(x, y), \quad \left|\frac{\partial^2 f(x, y; \alpha)}{\partial \alpha^2}\right| < F_2(x, y),$$

$$\left|\frac{\partial^3 f(x, y; \alpha)}{\partial \alpha^3}\right| < F_3(x, y)$$

astfel ca

$$\int_{-\infty}^{+\infty} F_3(x, y) f(x, y; \alpha)\, dy < M,$$

M. fiind independent de α.

III. *Funcția* $f(x, y; \alpha)>0$ *este mărginită în tot planul, pentru oricare* $\alpha\in\Delta$.

IV. *Există trei numere reale* h, k, l *astfel că, pentru oricare* $\alpha\in\Delta$, *integralele de mai jos există și avem*

$$\int_{-\infty}^{+\infty}\left[\frac{\partial \ln f(x, y; \alpha)}{\partial\alpha}\right]^2 f(x, y; \alpha)\,dy>h^2>0,$$

$$\int_{-\infty}^{+\infty}\left[\frac{\partial^2 \ln f(x, y; \alpha)}{\partial\alpha^2}\right] f(x, y; \alpha)\,dy>k^2>0,$$

$$\int_{-\infty}^{+\infty} F_3^2(x, y, f(x, y; \alpha)\,dy>l^2>0.$$

V. *Avem*

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} M\left[\frac{\partial \ln f(x_{i-1}, x_i; \alpha)}{\partial\alpha}\right]^2=C_1,$$

$$\lim_{n\to\infty}\frac{1}{n} M\left[\sum_{i=1}^{n} F_3(x_{i-1}, x_i)^2\right]=C_2,$$

unde C_1 și C_2 sînt niște constante.

Fie α_0 adevărata valoare a lui α.

Presupunînd că $\alpha_0\in\Delta$, avem, dezvoltînd în serie în raport cu α,

$$\frac{\partial \ln f(x, y; \alpha)}{\partial\alpha}=\left[\frac{\partial \ln f(x, y; \alpha)}{\partial\alpha}\right]_{\alpha=\alpha_0}+(\alpha-\alpha_0)\left[\frac{\partial^2 \ln f(x, y; \alpha)}{\partial\alpha^2}\right]_{\alpha=\alpha_0}+$$
$$+\frac{1}{2}\theta(\alpha-\alpha_0)^2 F_3(x, y),$$

prin θ însemnînd un număr real, astfel ca $|\theta|<1$.

Ținînd seamă de această formulă, ecuația verosimilității maxime (XX.19,9) devine

$$\sum_{i=1}^{n}\left[\frac{\partial \ln f(x_{i-1}, x_i; \alpha)}{\partial\alpha}\right]_{\alpha=\alpha_0}+(\alpha-\alpha_0)\sum_{i=1}^{n}\left[\frac{\partial^2 \ln f(x_{i-1}, x_i; \alpha)}{\partial\alpha^2}\right]_{\alpha=\alpha_0}+$$
$$+\frac{1}{2}\theta(\alpha-\alpha_0)^2\sum_{i=1}^{n} F_3(x_{i-1}, x_i)=0. \qquad \text{(XX.19,10)}$$

Să notăm, întocmai ca și în cazul variabilelor independente tratate de H. Cramér

$$B_0 = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\partial \ln f(x_{i-1}, x_i; \alpha)}{\partial \alpha}\right]_{\alpha=\alpha_0},$$

$$B_1 = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\partial^2 \ln f(x_{i-1}, x_i; \alpha)}{\partial \alpha^2}\right]_{\alpha=\alpha_0},$$

$$B_2 = \frac{1}{n}\sum_{i=1}^{n} F_3(x_{i-1}, x_i).$$

Ecuația verosimilității maxime se poate scrie

$$B_0 + B_1(\alpha-\alpha_0) + \frac{1}{2}\theta B_2(\alpha-\alpha_0)^2 = 0. \qquad \text{(XX.19,11)}$$

Să considerăm variabila ξ, dată de (XX.19,8), în punctul $\alpha=\alpha_0$. Din faptul că pentru $n\to\infty$, ξ tinde către legea normală, rezultă că B_0 tinde în probabilitate către zero, cînd $n\to\infty$, ținînd bine-înțeles seamă și de condiția III.

Să considerăm variabila

$$\eta = \frac{\sum_{i=1}^{n}\frac{\partial^2 \ln f(x_{i-1}, x_i; \alpha)}{\partial \alpha^2}}{\sqrt{\sum_{i=1}^{n} M\left[\frac{\partial^2 \ln f(x_{i-1}, x_i; \alpha)}{\partial \alpha^2}\right]^2}}.$$

Variabilele $\dfrac{\partial^2 \ln f(x_{i-1}, x_i; \alpha)}{\partial \alpha^2}$ reprezintă un lanț Markov, simplu și constant, formînd o singură mulțime finală. În baza condițiilor IV rezultă că η urmează la limită o distribuție normală. Același lucru se poate afirma și despre variabila

$$\zeta = \frac{\sum_{i=1}^{n} F_3(x_{i-1}, x_i)}{\sqrt{M\left[\sum_{i=1}^{n} F_3(x_{i-1}, x_i)\right]^2}}.$$

Rezultă atunci, avînd în vedere condițiile IV și V, că B_1 și B_2 tind în probabilitate către valorile finite C_1 și C_2.

Demonstrația urmează acum ca în cazul variabilelor independente. Oricît de mici ar fi numerele pozitive δ şi ε, putem găsi un N, astfel ca pentru $n>N$, să avem

$$P_1=P(|B_0|\geqq\delta^2)<\frac{\varepsilon}{3},$$

$$P_2=P\left(B_1\geqq\frac{1}{2}C_1\right)<\frac{\varepsilon}{3},$$

$$P_3=P(B_2\geqq 2C_2)<\frac{\varepsilon}{3}.$$

După inegalitatea lui Boole, probabilitatea ca punctul de selecție $(x_1, x_2, \ldots, x_n)$ să satisfacă inegalitățile

$$|B_0|<\delta^2, \quad B_1<\frac{1}{2}C_1, \quad B_2<2C_2 \qquad \text{(XX.19,12)}$$

este mai mare decît $1-\varepsilon$.

Pentru $\alpha=\alpha_0\pm\delta$, membrul intîi al ecuației (XX.19,11) devine

$$B_0\pm B_1\delta+\frac{1}{2}\theta B_2\delta^2.$$

In orice punct pentru care inegalitățile (XX.19,12) sînt îndeplinite, avem

$$\left|B_0\pm\frac{1}{2}\theta B_2\delta^2\right|<\delta^2(1+C_2).$$

Deci, dacă vom lua

$$\delta^2(1+C_2)<B_1\delta,$$

ceea ce revine la

$$\delta<\frac{B_1}{1+C_2}, \qquad \text{(XX.19,13)}$$

semnul expresiei $B_0\pm B_1\delta\pm\frac{1}{2}\theta B_2\delta^2$ depinde de semnul expresiei $\pm B_1\delta$. Prin urmare, putem alege pe δ satisfăcînd inegalitatea (XX.19,13) astfel ca, pentru $\alpha=\alpha_0\pm\delta$, membrul întîi al ecuației (XX.19,11) să ia valori egale şi de semn contrar.

Scanned with OKEN Scanner

De unde rezultă că, cu o probabilitate mai mare decît $1-\varepsilon$, ecuația verosimilității maxime (XX.19,13) admite o rădăcină cuprinsă între $\alpha_0-\delta$ și $\alpha_0+\delta$. Există deci o soluție a ecuației verosimilității care tinde în probabilitate către α_0.

Pe de altă parte, din (XX.19,11) rezultă

$$\alpha-\alpha_0=\frac{B_0}{-B_1-\frac{1}{2}\theta(\alpha-\alpha_0)B_2}.$$

Această relație se mai poate scrie sub forma

$$\frac{nC_1(\alpha-\alpha_0)}{\sqrt{\sum_{i=1}^{n}\left[\frac{\partial \ln f(x_{i-1},\, x_i;\, \alpha)}{\partial\alpha}\right]^2}}=\frac{\xi}{-\frac{B_1}{C_1}-\frac{1}{2C_1}\theta(\alpha-\alpha_0)B_0}.$$

Am văzut că, pentru $n\to\infty$, ξ tinde către distribuția normală, iar numitorul fracției din membrul al doilea al acestei egalități tinde în probabilitate către 1. Rezultă că fracția din membrul al doilea al egalității de mai sus tinde, pentru $n\to\infty$, către legea normală $N(0,\,1)$.

Ținînd seamă de acest rezultat și de condiția V, rezultă că membrul întîi al egalității considerate, care se poate scrie asimptotic

$$n\sqrt{C_1}(\alpha-\alpha_0)$$

tinde, pentru $n\to\infty$, către legea normală $N(0,\,1)$.

Rezultă atunci că

$$\lim_{n\to\infty}D^2(\alpha)=\frac{1}{\lim\limits_{n\to\infty}C_1 n}=\frac{1}{\lim\limits_{n\to\infty}\sum_{i=1}^{n}M\left[\frac{\partial \ln f(x_{i-1},\, x_i;\, \alpha)}{\partial\alpha}\right]^2}.$$

Minimul dispersiei funcțiilor de estimație este atins și deci putem în condițiile arătate să afirmăm că soluția ecuației verosimilității maxime admite o soluție asimptotic eficientă.

§ 20. Cazul lanțurilor cu legături complete

Sîntem într-un caz de mult mai mare generalitate cînd avem de estimat unul sau mai mulți parametri ai unui lanț cu legături complete.

In acest caz, dacă vom considera o selecție de mărime n

$$x_1, \quad x_2, \ldots, x_n \tag{XX.20,1}$$

păstrînd notațiile folosite pînă acum, avem

$$P(x_1, x_2, \ldots, x_n; \alpha) =$$

$$= f_1(x_1; \alpha) f_2(x_1, x_2; \alpha) f_3(x_1, x_2, x_3; \alpha) \ldots f_n(x_1, x_2, \ldots x_n; \alpha). \tag{XX.20,2}$$

Notăm cu $f_k(x_1, x_2, \ldots, x_k; \alpha)$ probabilitatea fundamentală de a avea pentru variabila de ordinul k din selecție o valoare cuprinsă între x_k și $x_k + dx_k$, α fiind un parametru care urmează să fie estimat.

Avem

$$f_k(x_1, x_2, \ldots, x_k; \alpha) \geqq 0, \int_{-\infty}^{+\infty} f(x_1, x_2, \ldots, x_k; \alpha)\, dx_k = 1. \tag{XX.20,3}$$

Fie $\bar{\alpha}(x_1, x_2, \ldots, x_k)$ o funcție de estimație pentru α. In același mod ca și pentru lanțul Markov, se arată că

$$D^2(\bar{\alpha}) \geqq \frac{1}{\sum_{k=1}^{n} M\left[\frac{\partial \ln f(x_1, x_2, \ldots, x_k; \alpha)}{\partial \alpha}\right]^2}; \tag{XX.20,4}$$

Putem să ne punem problema determinării condițiilor în care există o soluție a ecuației verosimilității maxime care să fie asimptotic normală și eficientă. Pentru aceasta, trebuie să determinăm cazurile cînd sumele de variabile, care se succed după legea unui lanț cu legături complete, tind la limită către distribuția normală. Condiții suficiente pentru existența legii normale ca limită pentru sumele de variabile complet înlănțuite au fost date în mai multe lucrări. Cele mai cuprinzătoare au fost găsite însă de G. Ciucu: *Proprietăți ergodice ale unor lanțuri cu legături complete* (Buletinul științific al Secției de Matematică al Academiei R. P. R. nr. 4, 1957).

§ 21. Cazul studiat de F. Chartier

In lucrarea sa *„L'estimation statistique dans le cas d'observations non indépendantes"* (Teză prezentată la Institutul de Statistică al Universității din Paris, 1947). F. Chartier a considerat pentru distribuția teoretică următoarea lege de probabilitate

$$f(x_{i-1}, x_i; \theta, \sigma, r) = \frac{1}{\sqrt{2\pi}\sigma\sqrt{1-r^2}} e^{-\frac{[(x_i-\theta)-r(x_{i-1}-\theta)]^2}{2\sigma^2(1-r^2)}}, \quad \text{(XX.21,1)}$$

adică un lanț Markov simplu și constant, în care intră trei parametri de estimat θ, σ, r.

Faptul că sumele de variabile înlănțuite după legea (XX.21,1) urmează la limită legea normală rezultă din teoria generală expusă la § 15.

Maximul de verosimilitate ne conduce la sistemul de ecuații

$$\sum_{i=1}^{n} \frac{\xi_i}{\sigma(1-r^2)} + r\sum_{i=1}^{n} \frac{\xi_{i-1}+\xi_i}{\sigma(1-r^2)} + r^2\sum_{i=1}^{n} \frac{\xi_i}{\sigma(1-r^2)} = 0,$$

$$-\frac{n}{\sigma} + \sum_{i=1}^{n} \frac{\xi_i^2}{\sigma(1-r^2)} - 2r\sum_{i=1}^{n} \frac{\xi_{i-1}\xi_i}{\sigma(1-r^2)} + r^2\sum_{i=2}^{n-1} \frac{\xi_i^2}{\sigma(1-r^2)} = 0,$$

$$\frac{(n-1)r}{1-r^2} + \sum_{i=1}^{n} \frac{(\xi_i - r\xi_{i-1})\xi_{i-1}}{1-r^2} - r\sum_{i=1}^{n} \frac{\xi_i - r\xi_{i-1}}{(1-r^2)^2} = 0, \quad \text{(XX.21,2)}$$

unde s-a notat

$$\xi_i = \frac{x_i - \theta}{\sigma}.$$

Dacă se presupune că σ și r sînt cunoscuți, din prima din ecuațiile de mai sus, corespunzătoare egalării cu zero a derivatei funcției de verosimilitate în raport cu θ, rezultă următoarea funcție de estimație pentru θ

$$\bar{\theta} = \frac{\sum_{i=1}^{n} x_i + r\left(\sum_{i=1}^{n-1} x_i + \sum_{i=2}^{n} x_i\right) + r^2\sum_{i=2}^{n-1} x_i}{n + 2(n-1)r + (n-2)r^2}.$$

$\bar{\theta}$ este o variabilă normală, avînd dispersia egală cu

$$D^2(\bar{\theta}) = \frac{1-r}{1+r} \cdot \frac{n}{\sigma^2}. \qquad \text{(XX.21,3)}$$

După cum se vede, există o pierdere de informație dacă r este pozitiv, cîştig în cazul contrar.

Din teorema pe care am enunțat-o la § 15 rezultă însă şi mai mult şi anume că $\bar{\theta}$ este o funcție de estimație eficientă a lui θ.

Intr-adevăr, un calcul simplu ne arată că

$$\sum_{i=1}^{n} M\left[\frac{\partial \ln f(x_{i-1}, x_i, \theta)}{\partial \theta}\right]^2 = \frac{(1-r)\,n}{(1+r)\,\sigma}.$$

Comparînd această egalitate cu (XX.21,3), rezultă proprietatea menționată.

Rezolvarea sistemului de ecuații (XX.21,2) conduce la determinarea unor funcții de estimație pentru r şi σ, asupra cărora nu mai insistăm. Este interesant că F. Chartier a aplicat rezultatele sale la o serie de o sută observații stelare făcute la Observatorul din Paris şi a constatat că ele se apropie mai mult de realitate decît datele teoretice rezultate din aplicarea teoriei clasice a erorilor de observație, bazată pe independența probelor.

Capitolul XXI

Procedee de estimare a funcției de distribuție

§ 1. Teoreme de convergență a frecvenței către probabilitate

1. La baza a numeroase aplicații ale statisticii matematice la fenomenele experimentale stă principiul după care *frecvența (relativă) experimentală a unui eveniment converge către frecvența sa teoretică.*

Potrivit teoremei lui Bernoulli, frecvența teoretică a unui eveniment cu probabilitate constantă p converge în probabilitate — după Cantelli chiar tare — către p.

Principiul citat şi teorema lui Bernoulli îngăduie deci să evaluăm direct probabilitatea p a evenimentului sau a valorii ce-i corespunde, printr-o frecvenţă experimentală.

Dacă o experienţă nu priveşte doar un singur eveniment ci sistemul de evenimente care corespunde unei variabile aleatorii, indicaţia dată de teorema lui Bernoulli este numai locală, ea poate folosi pentru analiza individuală a unora dintre evenimentele pe care le putem asocia unei valori a variabilei aleatorii.

Diferite cercetări au arătat însă că se poate extinde metoda care stă la baza teoremei lui Bernoulli pentru a obţine teoreme globale privind convergenţa funcţiei de frecvenţă $F_n(x)$ către funcţia de distribuţie $F(x)$ a variabilei considerate.

Presupunînd că în n probe posibile, independente, avem valorile

$$x_1, x_2, \ldots, x_n, \qquad \text{(XXI.1,1)}$$

funcţia de frecvenţă corespunzătoare este

$$F_n(a) = \frac{\nu_n(a)}{n},$$

unde $\nu_n(a)$ este numărul valorilor din şirul (XXI.1,1) mai mici decît a.

Funcţia de distribuţie este, potrivit definiţiei,

$$F(a) = P(x < a).$$

Teorema lui Glivenko afirmă că $F_n(a)$ converge tare către $F(a)$. Este deci posibil a aplica şi pentru funcţia de distribuţie principiul care leagă frecvenţa empirică de cea teoretică şi prin aceasta de distribuţia teoretică $F(a)$.

Prima teoremă asimptotică care evaluează de aproape distanţa între $F_n(a)$ şi $F(a)$ a fost dată de A. N. Kolmogorov, după care a urmat teorema lui Smirnov. Alte teoreme asimptotice pentru distanţa relativă $\frac{F_n(a)-F(a)}{F(a)}$ cînd $F(a) \geqq \varepsilon$ şi $\frac{F_n(a)-F(a)}{1-F(a)}$ pentru $1-F(a) \geqq \varepsilon$ au fost date de Renyi, Gnedenko şi alţii.

Legi mai precise, în care alături de termenul asimptotic apar şi termeni depinzînd de n şi fac posibilă o mai exactă evaluare a distanţei, au fost date de Gnedenko şi alţii.

Ceea ce rămîne a fi încă studiat pentru a da un fundament sistematic criteriilor de estimaţie a'e funcţiei $F(a)$ pe baza funcţiei de frecvenţă experiment lă $F_n^{(e)}(a)$ este posibilitatea estimării func-

ției $F_n(a)$ prin $F_n^{(e)}(a)$, această din urmă funcție fiind construită pe baza unor probe experimentale efective, așa cum $F_n(a)$ este construită pe baza sistemelor de probe teoretice posibile.

2. **Teorema lui Bernoulli.** *Probabilitatea teoretică a frecvenței* $\frac{\nu}{n}$ *a evenimentului A în n probe independente este* $C_n^\nu p^\nu (1-p)^{n-\nu}$, *dacă p este probabilitatea evenimentului A în fiecare probă.*

Rezultă că

$$P\left(\left|\frac{\nu}{n}-p\right|<k\frac{\sqrt{pq}}{\sqrt{n}}\right)=\sum_{|\lambda|<k} C_n^{np+\lambda\sqrt{npq}_p\, np_-+\lambda\sqrt{npq}_q\, nq-\lambda\sqrt{npq}},$$

unde

$$\lambda=\frac{\nu-np}{\sqrt{npq}}.$$

O evaluare simplă a acestei probabilități se poate efectua în modul următor

Funcția caracteristică a variabilei $\lambda=\frac{\nu-np}{\sqrt{npq}}$ este

$$\varphi_n(\tau)=\left(pe^{i\frac{\tau q}{\sqrt{npq}}}+qe^{-i\frac{\tau p}{\sqrt{npq}}}\right)^n.$$

Prin urmare, dezvoltînd în serie expresia din paranteză,

$$\varphi_n(\tau)=\left(1-\frac{1}{2}\cdot\frac{\tau^2}{n}+a\frac{\tau^3}{n\sqrt{n}}+b\frac{\tau^4}{n^2}+\ldots\right)^n,$$

deci

$$\varphi_n(\tau)=e^{-\frac{1}{2}\tau^2}\cdot e^{a\frac{\tau^3}{\sqrt{n}}+b\frac{\tau^4}{n}+O\left(\frac{1}{n}\right)},$$

unde a și b se calculează fără greutate în funcție de p și q.

Aceasta ne dă încă

$$\varphi_n(\tau)=e^{-\frac{1}{2}\tau^2}\left(1+a\frac{\tau^3}{\sqrt{n}}\right)+b\frac{\tau^4}{n}+\frac{1}{2}a^2\frac{\tau^6}{n}+O\left(\frac{1}{n}\right).$$

Scanned with OKEN Scanner

Funcția de distribuție corespunzătoare este dată, cu aproximațiile indicate de expresia precedentă, de integrala

$$F_n(\alpha)=\int_{-\infty}^{a}\rho_n(u)\,du,$$

cu

$$\rho_n(u)=\frac{1}{2\pi}\int_{-\infty}^{+\infty}e^{-i\tau u}\varphi_n(\tau)\,d\tau=\frac{1}{2\pi}\int_{-\infty}^{+\infty}e^{-i\tau u}e^{-\frac{1}{2}\tau^2}\,d\tau+$$

$$+\frac{1}{2\pi}\cdot\frac{a}{\sqrt{n}}\int_{-\infty}^{+\infty}e^{-i\tau u}\cdot e^{-\frac{1}{2}\tau^2}\cdot\tau^3\,d\tau+\frac{1}{2\pi n}\left(b\int_{-\infty}^{+\infty}e^{-i\tau u}\cdot e^{-\frac{1}{2}\tau^2}\cdot\tau^4\,d\tau+\right.$$

$$\left.+\frac{1}{2}a^2\int_{-\infty}^{+\infty}e^{-i\tau u}\cdot e^{-\frac{1}{2}\tau^2}\cdot\tau^6\,d\tau\right)+O\left(\frac{1}{n}\right)\cdot$$

Adică, efectuînd integralele în modul obișnuit,

$$\rho_n(u)=\frac{1}{\sqrt{2\pi}}e^{-\frac{u^2}{2}}+\frac{1}{\sqrt{2\pi}\sqrt{n}}e^{-\frac{u^2}{2}}(Au+Bu^3)+$$

$$+\frac{1}{2\pi n}\cdot\frac{1}{\sqrt{2\pi n}}e^{-\frac{u^2}{2}}(C+Du^2+Eu^4+Fu^6)+O\left(\frac{1}{n}\right),$$

A, B, C, D, E, F fiind independente de n și de u.
Deoarece probabilitatea inegalității $\lambda<\alpha$ este

$$F_n(\alpha)=\int_{-\infty}^{+a}\rho_n(u)\,du,$$

rezultă

$$P\left(\frac{\nu}{n}-p\leqslant\frac{\alpha\sqrt{pq}}{\sqrt{n}}\right)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+a}e^{-\frac{x^2}{2}}dx+\frac{1}{\sqrt{n}}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+a}(Ax+Bx^3)e^{-\frac{x^2}{2}}dx+$$

$$+\frac{1}{n}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+a}(C+Dx^2+Ex^4+Fx^6)\,e^{-\frac{x^2}{2}}dx+O\left(\frac{1}{n}\right),\qquad\text{(XXI.1,2)}$$

iar probabilitatea ca $|\lambda|<\alpha$ fiind

$$F_n(\alpha)-F_n(-\alpha)=\int_{-a}^{+a}\rho_n(u)\,du,$$

găsim

$$P\left(\left|\frac{\nu}{n}-p\right| \leqslant \frac{\alpha\sqrt{pq}}{\sqrt{n}}\right)=\frac{1}{\sqrt{2\pi}}\int_{-\alpha}^{+\alpha} e^{-\frac{x^2}{2}}dx+\frac{N(\alpha)}{n}e^{-\frac{\alpha^2}{2}}+O\left(\frac{1}{n}\right), \tag{XXI.1,3}$$

$N(\alpha)$ fiind un polinom în α.

Aceste două expresii ne dau criterii pentru aprecierea alegerii valorii p drept probabilitate şi anume criterii asimptotice (pentru $n\to\infty$), care sînt independente de p, şi criterii care depind de n şi, precum se vede lesne, şi de p.

Criteriile asimptotice sînt exprimate cu ajutorul legii normale.

3. *Generalizarea lui Poisson* a teoremei lui Bernoulli ne interesează în special dacă probabilitatea p_j ($j=1, 2, \ldots, n, \ldots$), presupusă variabilă cu proba, are numai o mică oscilaţie în jurul unei valori fixe p ($0<p\leqslant 1$)

$$p_j=p+\alpha_j, \quad |\alpha_j|<\varepsilon, \quad j=1, 2, \ldots, n, \ldots$$

In acest caz calculele vor diferi, dar nu esenţial, de cele efectuate în cazul precedent.

Să punem

$$n\pi'_n=p_1+p_2+\ldots+p_n,$$
$$n\beta_n=\alpha_1+\alpha_2+\ldots+\alpha_n$$

şi

$$\pi'_n=p+\beta_n, \quad \text{cu} \quad \beta_n<\varepsilon, \qquad 0\leqslant\pi'_n\leqslant 1.$$

Vom lua, de asemenea, un σ_n pentru care

$$n\sigma_n(1-\sigma_n)=\sum_{j=1}^{n}p_j(1-p_j),$$

expresie care reprezintă dispersia variabilei sumă

$$n\sigma_n(1-\sigma_n)=npq+(1-2p)n\beta_n+n\gamma_n,$$

unde $n\gamma_n=\sum\alpha_j^2$.

Variabila normată a sumei este în acest caz $\lambda=\frac{\nu-n\pi_n}{\sqrt{n\sigma_n(1-\sigma_n)}}$, iar funcția caracteristică este

$$\varphi_n(\tau)=\prod_{j=1,\dots,n}\left(p_j e^{i\frac{q_j}{\sqrt{n\sigma_n(1-\sigma_n)}}}+q_j e^{-i\frac{p_j}{\sqrt{n\sigma_n(1-\sigma_n)}}}\right).$$

Deci

$$\ln\varphi_n(\tau)=-\frac{\tau^2}{2}+a_\varepsilon\frac{\tau^3}{\sqrt{n}}+b_\varepsilon\frac{\tau^4}{n}+O\left(\frac{1}{n}\right),$$

unde $a_\varepsilon\to a$, $b_\varepsilon\to b$ cînd $\varepsilon\to 0$. Dacă ε este independent de n, viteza de convergență diferă, în cazul Poisson, de cazul Bernoulli.

Dar dacă luăm $\varepsilon=\frac{1}{n}$, avem și în acest caz

$$\ln\varphi_n(\tau)=-\frac{\tau^2}{2}+a\frac{\tau^3}{\sqrt{n}}+b\frac{\tau^4}{n}+O\left(\frac{1}{n}\right),$$

deci viteza de convergență depinde de $\frac{1}{\sqrt{n}}$ și — dacă se poate spune, avem o supraviteză de convergență $\left(\text{în }\frac{1}{n}\right)$ — este aceeași ca în cazul Bernoulli.

4. Variabila f_n-f_m are funcția caracteristică

$$\varphi_{n,m}(\tau)=\left(pe^{i\frac{\tau}{n}}+q\right)^n\left(pe^{-i\frac{\tau}{m}}+q\right)^m, \qquad \text{(XXI.1,4)}$$

ceea ce dă

$$M[f_n-f_m]=M(f_n)-M(f_m)=0,$$

$$M[(f_n-f_m)^2]=\frac{n+m}{nm}pq.$$

Variabila normată $\frac{f_n-f_m}{\sqrt{\frac{n+m}{nm}pq}}$ are ca funcție caracteristică

$$\psi_{n,m}=\varphi_{n,m}\left(\frac{\tau}{\sqrt{\frac{n+m}{nm}pq}}\right).$$

Deoarece

$$\varphi_{n,m}(\tau)=\left(1+ip\frac{\tau}{n}-\frac{1}{2}p\frac{\tau^2}{n^2}-\frac{i}{6}p\frac{\tau^3}{n^3}+\ldots\right)^n\times$$
$$\times\left(1-ip\frac{\tau}{m}-\frac{1}{2}p\frac{\tau^2}{m^2}+\frac{i}{6}p\frac{\tau^3}{m^3}+\ldots\right)^m,$$

avem

$$\ln\varphi_{n,m}=-\frac{1}{2}pq\left(\frac{1}{m}+\frac{1}{n}\right)\tau^2-\frac{i}{6}pq(2p-1)\left(\frac{1}{n^2}-\frac{1}{m^3}\right)\tau^3+\ldots$$

Rezultă atunci

$$\ln\psi_{n,m}(\tau)=-\frac{\tau^2}{2}+i\frac{2p-1}{\sqrt{pq}}\frac{\frac{1}{n}-\frac{1}{m}}{\sqrt{\frac{1}{n}+\frac{1}{m}}}\frac{\tau^3}{6}+\ldots \qquad \text{(XXI.1,5)}$$

$$\lim_{n,m\to\infty}\psi_{n,m}(\tau)=e^{-\frac{\tau^2}{2}}. \qquad \text{(XXI.1,6)}$$

Distribuția asimptotică este cea normală, așa cum era de așteptat. Formula (XXI.1,5) ne dă însă și o indicație asupra modului cum este realizată această distribuție asimptotică în funcție de m și n.

5. Teorema lui Cantelli. *Convergența frecvenței* $\frac{\nu}{n}$ *către probabilitatea constantă* p *a evenimentului* A *are loc nu numai în probabilitate*, cum a arătat teorema lui Bernoulli și relația mai precisă (XXI.1,2), ci, precum a arătat Cantelli, *această convergență este tare sau aproape sigură*. Iată teorema lui Cantelli

Fiind date ε *și* η, *există* $N(\varepsilon,\eta)$ *astfel că*

$$P\{|f_n-p|<\varepsilon,\ |f_{n+1}-p|<\varepsilon,\ldots\}>1-\eta \qquad \text{(XXI.1,7)}$$

dacă $n>N(\varepsilon,n)$, f_n *fiind frecvența în* n *probe* $\left(f_n\equiv\frac{\nu}{n}\right)$.

In adevăr, potrivit inegalității generalizate a lui Cebîșev,

$$P\{|f_n-p|<\varepsilon\}>1-\frac{M_4^{(n)}}{\varepsilon^4}>1-\frac{1}{\varepsilon^4n^2}, \qquad \text{(XXI.1,8)}$$

deoarece avem pentru momentul de ordinul patru următoarea inegalitate

$$M_4^{(n)} = M\left[\left(\frac{\nu}{n} - p\right)^4\right] = \frac{1}{n^2}\left[3p^2q^2 - \frac{6p^2q^2}{n} + \frac{pq}{n}\right] < \frac{1}{n^2}.$$

Dacă E este spațiul evenimentelor unei singure probe, să considerăm spațiul $E^\infty = E \times E \times \ldots$ al evenimentelor succesiunilor de probe.

Fie A_n evenimentul pentru care $|f_n - p| < \varepsilon$. După (XXI.1,8) avem

$$P(A_n) > 1 - \frac{1}{\varepsilon^4 n^2}.$$

Primul membru din (XXI.1,7) este $P(A_n \cap A_{n+1} \cap \ldots)$, iar după teorema lui Boole

$$P\left(\bigcap_{m \geqq n} A_m\right) > 1 - \sum_{m \geqq n} P(\cap A_m) > 1 - \sum_{m \geqq n} \frac{1}{\varepsilon^2 m^2}.$$

Deoarece seria $\sum \frac{1}{m^2}$ este convergentă, putem alege un N suficient de mare, astfel ca, pentru $n \geqq N$, restul $\sum_{m \geqq n} \frac{1}{m^2} < \eta\varepsilon^2$ și astfel inegalitatea (XXI.1,7) este realizată.

Prin urmare, frecvența considerată ca variabilă aleatorie converge tare către p.

De aici rezultă imediat că mulțimea M din E^∞, pe care frecvența converge către p, are probabilitatea egală cu unu

$$P(M) = 1, \qquad \text{(XXI.1,9)}$$

ceea ce exprimă faptul că frecvența converge aproape sigur către probabilitate.

§ 2. Convergența funcției $F_n(x)$ către $F(x)$

1. Teorema lui Glivenko. In loc de frecvența unei singure valori și probabilitatea ei, să urmărim frecvențele valorilor aceleiași variabile aleatorii în comparație cu probabilitățile respective. Această operație trebuie considerată global, așa cum se vede în următoarea teoremă a lui Glivenko.

Dacă funcția de distribuție $F(x)$ a variabilei aleatorii considerate $x(\xi)$ este continuă și notăm cu

$$d_n = \sup_a |F_n(a) - F(a)|,$$

distanța dintre funcția de distribuție $F(x)$ și frecvența teoretică $F_n(a)$ (=numărul valorilor mai mici decît ale variabilei $x = x(\xi)$ în n probe), *atunci d_n converge aproape sigur către zero cînd $n \to \infty$.*

In demonstrație intervine în mod esențial *teorema lui Cantelli.*

Fie $a_{r,k}$ ($k = 0, 1, 2, \ldots, r$; $r = 0, \pm 1, \pm 2, \ldots$) un dublu sistem de numere reale, crescător, cînd r este fix, odată cu k și astfel că, dîndu-se un ε fix și un a oarecare, să existe un r și un k astfel ca

$$0 \leqq a_{r,k} - a < \frac{\varepsilon}{2}, \quad 0 < a - a_{r,k+1} < \frac{\varepsilon}{2}, \qquad \text{(XXI.2,1)}$$

$F(a_{r,k})$ este probabilitatea evenimentului pentru care $x < a_{r,k}$.

Frecvența $F_n(a_{r,k})$ este o variabilă aleatorie care, potrivit teoremei lui Cantelli, converge aproape sigur către $F(a_{r,k})$. Dacă numim cu $M_{r,k}$ mulțimea din E^∞ pe care are loc această convergență, potrivit cu (XXI.1,9), vom avea

$$P(M_{r,k}) = 1.$$

Pe de altă parte, mulțimea pentru care $\sup_{k=1,2,\ldots,r} |F_n(a_{r,k}) - F(a_{r,k})| \to 0$ este

$$M_r = M_{r,1} \cap M_{r,2} \cap \ldots \cap M_{r,r}.$$

Potrivit propoziției auxiliare de la sfîrșitul paragrafului precedent,

$$P(M_r) = 1$$

și, în virtutea aceleiași propoziții, dacă

$$M = \bigcup_{r=1,2,\ldots,\infty} M_r,$$

avem

$$P(M) = 1. \qquad \text{(XXI.2,2)}$$

Cum M este mulțimea pe care $\sup |F_n(a_{r,k}) - F(a_{r,k})| \to 0$ pentru orice k și orice r, (XXI.2,2) va arăta că această convergență este aproape sigură.

Fie acum mulțimea B pentru care $\sup |F_r(a)-F(a)| \to 0$ pentru orice a.

Alegem, în conformitate cu (XXI.2,1), un $a_{r,k}$ vecin după voie cu a.

Avem atunci $F_n(a_{r,k})-F(a_{r,k})=F_n(a_{r,k})-F_n(x)+F_n(x)-F(x)+$

$+F(x)-F(a_{r,k})$.

Deoarece

$$F_n(a_{r,k+1})-F_n(x)<F_n(a_{r,k+1})-F_n(a_{r,k}),$$

rezultă că

$$\sup |F_n(a_{r,k})-F(a_{r,k})| \leqslant \sup |F_n(a_{r,k})-F_n(a_{r,k+1})| +$$
$$+\sup |F_n(1)-F(1)|+\sup |F(a)-F(a_{r,k})|,$$

deci $B \subset M$.

Tot așa se arată că $M \subset B$, deci $B=M$. Prin urmare, mulțimea B de convergență tare a variabilei $\sup |F_n(a)-F(a)|$ către zero are probabilitatea 1.

Teorema este demonstrată.

Teorema lui Glivenko este valabilă și pentru cazul general al unei funcții de distribuție cu discontinuități, dacă considerăm în expresia distanței

$$d_n=\sup_a |F_n(a)-F(a)|$$

numai punctele a de continuitate ale lui $F(a)$. Demonstrația nu suferă altă modificare decît alegerea sistemului $a_{r,k}$ astfel ca să nu cuprindă punctele de discontinuitate.

2. Observație importantă. Este necesar să semnalăm caracterul teoretic al funcției $F_n(a)$, adică al frecvenței evenimentului $x<a$ în n probe. Acest caracter este implicat în faptul că am aplicat acestei frecvențe teorema lui Cantelli. *Cînd se vorbește de frecvența empirică, aceasta se face numai în baza unui principiu care stabilește corespondența între cele două frecvențe.*

§ 3. Cîteva teoreme generale de limită

1. *Generalități și notații.* Cîteva noi teoreme generale de limită datorite lui A. Wald, K. L. Chung, Erdös și Kaç, M. Donsker și Renyi, ale cărui procedee le urmăm în acest paragraf deaproape, sînt

indispensabile pentru lămurirea comportării asimptotice a diferenţei $F_n(a)-F(a)$ sau a diferenţei relativizate $\frac{F_n(a)-F(a}{F(a)}$.

La sfîrşitul acestui paragraf, dăm şi cele două propoziţii ale lui Cramér, care au folosit în capitolul precedent, precum şi propoziţiile de caracter general, care au folosit în demonstraţia teoremei lui Glivenko.

Să considerăm un dublu sistem de variabile aleatorii

$$x_{n,1},\ x_{n,2},\ldots,\ x_{n,r_n}, \tag{XXI.3,1}$$

pentru $n=1,2,\ldots,\ \infty$ şi cu $r_n\to\infty$ cînd $n\to\infty$, şi să presupunem că toate au valoarea medie nulă şi dispersia finită.

Presupunem că variabilele $x_{n,1},\ x_{n,2},\ldots,\ x_{n,r_n}$ pentru un n dat sînt independente între ele şi verifică o condiţie de tip Lindeberg. Folosind notaţiile

$$s_{n,h}=\sum_{j=1,2,\ldots,h} x_{n,j},\quad s_n=s_{n,r_n}$$

$$\mu_{n,j}=M[x_{n,j}^2],\quad \mu_n^2=\sum\mu_{n,j}^2, \tag{XXI.3,2}$$

condiţia Lindeberg se exprimă prin

$$\lim_{n\to\infty}\frac{1}{\mu_n^2}\sum_{j=1}^{r_n}\int a^2dF_{n,j}(a)\to 0\qquad a>\tau\mu_n \tag{XXI.3,3}$$

pentru orice $\tau>0$.

2. In aceste condiţii, demonstrăm, urmînd un procedeu al lui Rényi, mai multe teoreme a căror semnificaţie pentru problemele noastre este imediată.

Teorema 1.

$$\lim P\ (\sup_{1<h\leq r_n} s_{n,h}<a\,\mu_n)=\begin{cases}\sqrt{\frac{2}{\pi}}\int\limits_0^a e^{-\frac{u^2}{2}}du & a>0\\ 0 & a\leq 0.\end{cases}$$

In vederea demonstrării acestei teoreme, să considerăm un întreg l şi cel mai mic întreg m_j $(j=1, 2, \ldots, l)$ pentru care

$$\sum_{h=1}^{m_j} \mu_{n,\,h}^2 \geqq \frac{j}{k}\,\mu_n^2. \qquad \text{(XXI.3,4)}$$

Este evident că

$$1 \leqq m_1 \leqq m_2 \leqq \ldots \leqq m_k = r_n.$$

Fie acum următoarele sume parţiale

$$\sigma_{n,\,1} = S_{n,\,m}, \quad \sigma_{n,\,2} = S_{n,\,m_2} - S_{n,\,m_1}, \ldots, \quad \sigma_{n,\,k_n-1} = S_{n,\,m_{k_n}} - S_{n,\,m_{k_n-1}}.$$

Fiecare dintre aceste sume parţiale, de pildă

$$\sigma_{n,\,j} = x_{n,\,m_{j-1}+1} + x_{n,\,m_{j-1}+2} + \ldots + x_{n,\,m_j},$$

este formată dintr-o succesiune de variabile care satisface, cum este aproape evident, condiţia lui Lindeberg.
Prin urmare, dacă

$$\mu_{n,\,j}^2 = M\,[\sigma_{n,\,j}^2], \qquad \text{(XXI.3,5)}$$

putem trage concluzia că

$$\lim_{n\to\infty} P\,(\sigma_{n,\,j} < a\mu_{n,\,j}) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-\frac{u^2}{2}}\,du. \qquad \text{(XXI.3,6)}$$

Dar variabilele $\sigma_{n,\,1}, \sigma_{n,\,2}, \ldots, \sigma_{n,\,k}$ sînt independente şi, ţinînd seamă de (XXI.3,4) şi (XXI.3,5),

$$\lim_{n\to\infty} \frac{\mu_{n,\,j}}{\mu_n} = \frac{1}{\sqrt{k}}.$$

Avem deci

$$\lim_{n\to\infty} P\left(\frac{\sigma_{n,\,1}}{\mu_n} < \frac{a_1}{\sqrt{k}},\ \frac{\sigma_{n,\,2}}{\mu_n} < \frac{a_2}{\sqrt{k}}, \ldots,\ \frac{\sigma_{n,\,k}}{\mu_n} < \frac{a_k}{\sqrt{k}}\right) =$$

$$= \frac{1}{(\sqrt{2\pi})^k} \int_{-\infty}^{a_1} \int_{-\infty}^{a_2} \ldots \int_{-\infty}^{a_k} e^{-\frac{1}{2}\sum_{h=1}^{k} u_h^2}\,du_1\,du_2 \ldots\,du_k.$$

Pentru acelaşi motiv

$$\lim P(\sigma_{n,1} < a\mu_n, \sigma_{n,1}+\sigma_{n,2} < a\mu_n, \ldots \sigma_{n,1}+\sigma_{n,2}+\ldots+\sigma_{n,k} < a\mu_n) =$$

$$= \frac{1}{(\sqrt{2\pi})^k} \int_{D_a} e^{-\frac{1}{2}\sum_{h=1}^{k} u_h^2} du_1\, du_2 \ldots du_k, \qquad \text{(XXI.3,7)}$$

domeniul D_a fiind

$$-\infty < u_1 < a\sqrt{k}, \quad -\infty < u_1+u_2 < a\sqrt{k}, \ldots, \quad -\infty < u_1 + u_2+\ldots+u_k < a\sqrt{k}. \qquad \text{(XXI.3,8)}$$

Observăm că expresia din membrul al doilea al egalităţii (XXI.3,7) este independentă de datele speciale ale problemei, prin urmare valoarea membrului întîi poate fi obţinută fie calculînd direct membrul al doilea, cu domeniul D_a dat de (XXI.3,8), fie calculînd chiar membrul întîi pentru valori speciale ale variabilelor $x_{n,h}$.

In prealabil să demonstrăm, însă, faptul intuitiv exprimat de egalitatea următoare

$$\lim P(\sigma_{n,1} < a\mu_n, \ \sigma_{n,1}+\sigma_{n,2} < a\mu_n, \ldots \sigma_{n,1}+\sigma_{n,2}+\ldots+ \sigma_{n,k} < a\mu_n) = P_n(a),$$

unde am pus, pentru prescurtarea scrisului

$$P_n(a) = P(s_{n,1} < a\mu_n, \ s_{n,2} < a\mu_n, \ldots, \ s_{n,r_n} < a\mu_n).$$

In acest scop, să punem

$$Q_{n,k}(a) = P(\sup_{1 \le j \le k} s_{n,m_j} < a\mu_n)$$

şi

$$\pi_{n,r}(a) = P(s_{n,r} \ge a\mu_n, \sup_{1 \le j \le r-1} s_{n,j} < a\mu_n).$$

Este vizibil că

$$\sum_{r=1}^{r_n} \pi_{n,r}(a) = 1 - P_n(a).$$

Fie j întregul pentru care $m_{j-1} < r \leq m_j$; introducem notațiile

$$\pi^{(1)}_{n,r}(a) = P\left(s_{n,2} \geq a\mu_n;\ \sup_{1 \leq j \leq r-1} s_{n,j} < a\mu_n;\ |s_{n,m_j} - s_{n,r}| \geq \varepsilon\mu_n\right)$$

și

$$\pi^{(2)}_{n,r}(a) = P\left(s_{n,2} \geq a\mu_n;\ \sup_{1 \leq j \leq r-1} s_{n,j} < a\mu_n;\ |s_{n,m_j} - s_{n,r}| < \varepsilon\mu_n\right),$$

ceea ce dă

$$\pi_{n,r}(a) = \pi^{(1)}_{n,r}(a) + \pi^{(2)}_{n,r}(a).$$

Inegalitatea lui Cebîșev conduce la

$$\pi^{(1)}_{n,r}(a) = \pi_{n,r}(a) \cdot P\left(s_{n,m_j} - s_{n,r}) \geq \varepsilon\mu_n\right) \leq \pi_{n,r}(a) \frac{\mu^2_{n,j}}{\varepsilon^2 \mu^2_n}.$$

Deci

$$\pi^{(1)}_{n,r}(a) \leq \pi_{n,r}(a) \frac{1+\theta}{\varepsilon^2 k},$$

de unde

$$1 - P_n(a) = \sum_{1,2,\ldots,r_n} \pi_{n,r}(a) \leq \frac{1+\theta}{\varepsilon^2 k} + \sum_{r=1,2,\ldots,r_n} \pi^{(2)}_{n,r}(a).$$

In acelaşi timp

$$1 - Q_{n,k}(x-\varepsilon) \geq \sum_{r=1,2,\ldots,r_n} \pi^{(2)}_{n,r}(a),$$

iar

$$s_{n,m_j} > (a-\varepsilon)\mu_n.$$

Prin urmare

$$1 - P_n(a) \leq \frac{1+\theta}{\varepsilon^2 k} + 1 - Q_{n,k}(a-\varepsilon).$$

Deci

$$Q_{n,k}(a-\varepsilon) - \frac{1+\theta}{\varepsilon^2 k} \leq P_n(a) \leq Q_{n,k}(a),$$

a doua parte a inegalității fiind evidentă.
Rezultă de aici egalitatea asimptotică cerută. Egalitatea (XXI.3,7) se mai poate scrie

$$\lim P_n(a) = \frac{1}{(\sqrt{2\pi})^k} \int_{D_a} e^{-\frac{1}{2}\sum_{j=1}^{k} u^2} du_1 \ldots du_k \qquad (a > 0).$$

In loc de a calcula direct membrul al doilea, putem calcula membrul întîi într-un caz particular.

Presupunem $x_{n,\,h}=\pm 1$, fiecare dintre aceste valori fiind luată cu probabilitatea $\frac{1}{2}$. Condiţiile cerute sînt îndeplinite: $M(x_{n,\,h})=0$, $M(x_{n,\,h}^2)=\mu_{n,\,h}^2=1$. Variabilele fiind mărginite uniform, condiţia lui Lindeberg este verificată.

Avem evident

$$P_n(a)=\sum_{\substack{-n<l<[a\sqrt{n}]\\ \nu\equiv n \pmod 2}}\left[C_{\frac{n+\nu}{2}}^{n}-C_{\frac{n-\nu}{2}+[a\sqrt{n}+1]}^{n}\right].$$

Aplicînd formula lui Stirling, găsim

$$\lim P_n(a)=\sqrt{\frac{2}{\pi}}\int_0^a e^{-\frac{u^2}{2}}\,du,$$

pentru $a>0$ şi

$$\lim P_n(a)=0$$

pentru $a\leqq 0$.

Teorema 1 este astfel demonstrată.

Teorema 2.

$$\lim_{n\to\infty}P(-b\mu_n\leqq\inf_{h=1,2,\ldots,r_n}s_{n,\,k}\leqq\sup_{h=1,2,\ldots,r_n}s_{n,\,h}<a\mu_n)=\lim P_n(a,\,b)=$$

$$=\frac{4}{\pi}\sum_{l=0}^{\infty}\frac{e^{-\frac{(2l+1)^2\pi^2}{2(a+b)^2}}\sin(2l+1)\pi\frac{a}{a+b}}{al+1},\qquad a\geqq 0,\quad b\geqq 0,$$

limita fiind nulă dacă unul dintre numerele a sau b este negativ.

Procedînd ca în teorema precedentă, găsim că $\lim_{n\to\infty}P_n(a,\,b)$ este o funcţie determinată de a şi b, care nu depinde de nici un alt element particular sistemelor $x_{n,\,k}$.

Dacă luăm aceleaşi variabile aleatorii $x_{n,k}=\pm 1$ cu probabilităţile respectiv egale cu $\frac{1}{2}$, deoarece $\mu_n^2=n$, avem

$$P(-b\sqrt{n}<s_{n,1}<a\sqrt{n},\ -b\sqrt{n}<s_{n,2}<a\sqrt{n},\ldots$$

$$\ldots,-b\sqrt{n}<s_{n,n}<a\sqrt{n})=\sum_{-\beta<l<\alpha}\left[\varphi_l+\sum_{r=1}^{\infty}(\varphi_{2r(\alpha+\beta)+l}-\varphi_{2r(\alpha+\beta)-2\beta-k}+\right.$$

$$\left.+\varphi_{-2r(\alpha+\beta)+k}-\varphi_{-2r(\alpha+\beta)+2\alpha-k})\right], \qquad \text{(XXI.3,9)}$$

unde $\alpha=[a\sqrt{n}]+1$, $\beta=[b\sqrt{n}+1]$ şi

$$\varphi_l=\begin{cases}\dfrac{1}{2^n}C_{\frac{n+l}{2}}^{n} & \text{pentru } k\equiv n \pmod 2\\ 0 & \text{pentru } k\not\equiv n \pmod 2.\end{cases}$$

Deoarece, după procedeul aplicat la teorema lui Bernoulli, avem

$$\lim_{n\to\infty}\frac{1}{2^n}\sum_{\frac{n}{2}-v\frac{\sqrt{n}}{2}<j<\frac{n}{2}+u\frac{\sqrt{n}}{2}}C_n^j=\frac{1}{\sqrt{2\pi}}\int_{-v}^{u}e^{-\frac{t^2}{2}}dt,$$

formula (XXI.3,9) devine

$$\lim P_n(a,\ b)=\frac{4}{\pi}\sum_{l=0}^{\infty}\frac{e^{-\frac{(2l+1)^2\pi^2}{2(a+b)^2}}\sin(2l+1)\pi\frac{a}{a+b}}{2l+1} \qquad \text{(XXI.3,10)}$$

pentru $a>0$, $b>0$.

Observaţie. Dacă $b=a$, această teoremă devine

$$\lim P(\sup_{r=1,2,\ldots,r_n}|s_{n,r}|<a\mu_n)=\frac{4}{\pi}\sum_{l=0}^{\infty}(-1)^l\frac{e^{-\frac{(2l+1)^2\pi^2}{8a^2}}}{2l+1},\qquad a>0, \qquad \text{(XXI.3,11)}$$

această limită fiind nulă pentru $a\leq 0$.

2. Vom expune în cele ce urmează două *Propoziţii ale lui H. Cramér.*

Propoziţia 1. *Fie x_n şi y_n două variabile aleatorii, ultima fiind infinitezimală cu $\frac{1}{n}$ în înţelesul că*

$$\lim_{n\to\infty}M[y_n]=0,\quad \lim_{n\to\infty}[(y_n-M(y_n))^2]=0.$$

Dacă funcţia de distribuţie $F_n(a)$ a variabilei x_n are ca limită $F(a)$ cînd $n\to\infty$ şi dacă numim cu $G_n(a)$ funcţia de distribuţie a variabilei x_n+y_n, atunci

$$\lim_{n\to\infty} G_n(a)=\lim_{n\to\infty} F_n(a)=F(a).$$

Pentru demonstraţie, vom presupune, fără să restrîngem generalitatea, că $M(y_n)=0$, deci $\lim M[y_n^2]=0$ cînd $n\to\infty$.
După inegalitatea lui Cebîşev, avem

$$P(|y_n|<\lambda)<\frac{M(y_n^2)}{\lambda^2}.$$

Există deci un întreg N astfel că

$$P(|y_n|<l)<\varepsilon,$$

dacă $n>N(\delta)$.
Să observăm că

$$\{x_n+y_n<a\}\subset\{x_n<a+\lambda\}\cup\{x_n\geqq a+\lambda\}.$$

Insă

$$x_n\geqq a+\lambda\subset\{y_n<-\lambda\}.$$

Prin urmare

$$G_n(a)=P(\{x_n+y_n<a\})\leqq P(\{x_n<a+\lambda\})+P(\{y_n<-\lambda\}).$$

In acelaşi mod găsim

$$G_n(a)\geqq P(\{x_n<a-\lambda\})-P(\{y_n>\lambda\}),$$

de unde

$$F_n(a-\lambda)-\varepsilon\leqq G_n(a)\leqq F_n(a+\lambda)+\varepsilon.$$

Dacă a este un punct de continuitate al lui $F(x)$, cînd facem $n\to\infty$ găsim

$$F(a)-\varepsilon\leqq\underline{\lim}\, G_n(a)<\overline{\lim}\, G_n(a)\leqq F(a)+\varepsilon,$$

oricare ar fi ε, deci

$$\underline{\lim}\, G_n(a)=\overline{\lim}\, G_n(a)=F(a).$$

Propoziţia este demonstrată.

Propoziția 2. *Fie x_n, y_n, z_n trei variabile aleatorii, primele două fiind independente între ele. Numind cu $F_n(a)$ și $G_n(a)$ funcțiile lor de distribuție, presupunem că există $F(a) = \lim F_n(a)$ și $G(a) = \lim G_n(a)$ în înțelesul obișnuit.*

Dacă z_n este asimptotic infinitezimală, variabilele $x_n + z_n$ și y_n sînt asimptotic independente.

Trebuie deci să stabilim că, presupunînd

$$\lim_{n\to\infty} M(z_n) = 0, \quad \lim_{n\to\infty} M[(z_n - M(z_n))^2] = 0,$$

vom avea

$$\lim_{n\to\infty} P\{x_n + z_n < a, \quad y_n < b\} = F(a) \cdot G(b).$$

Pentru aceasta, să observăm că se poate determina un $N(\lambda, \varepsilon)$ astfel ca, pentru $n > N(\lambda, \varepsilon)$ să avem

$$P(|y_n| < \lambda) < \varepsilon$$

și

$$P(x_n < a - \lambda, \ y_n < b) - \varepsilon \leq P(x_n + z_n < a, \ y_n < b) \leq P(x_n < a + \lambda, \ y_n < b) + \varepsilon.$$

Deoarece x_n și y_n sînt independente, relația precedentă se mai poate scrie

$$F_n(a - \lambda)\, G_n(b) - \varepsilon \leq P(x_n + z_n < a, \ y_n < b) \leq F_n(a + \lambda)\, G_n(b) + \varepsilon,$$

ceea ce dă la limită relația căutată, deci independența.

3. *Propozițiile auxiliare* care urmează prezintă interes, în special în legătură cu demonstrația teoremei lui Glivenko.

a) *$A_n (n \in N)$ fiind o infinitate de evenimente ale unui cîmp, în relațiile de incluziune $A_{n+1} \subset A_n$ $(n \in N)$, punînd*

$$A = \bigcap_{n\in N} A_n,$$

avem

$$p(A) = \lim P(A_n).$$

Demonstrația este imediată.

b) *Dacă evenimentele $A_1, A_2, \ldots$ au toate probabilitatea 1, atunci*

$$P\left(\bigcap_{j\in N} A_j\right) = 1.$$

Demonstrăm propoziţia mai întîi pentru cazul $A_{n+1}=A_{n+2}=\ldots=E$, deci

$$\bigcap_{j\in N} A_j=\bigcap_{j=1}^{n} A_j.$$

Considerăm atunci cazul unui număr finit de evenimente. Vom începe prin a verifica teorema pentru două evenimente A_1 şi A_2. Numind cu CA_1 şi CA_2 complementarele lor, avem

$$A_1=A_1\cap A_2+A_1\cap CA_2,$$
$$A_2=A_1\cap A_2+A_2\cap CA_1,$$
$$A_1\cup A_2=A_1\cap A_2+A_1\cap CA_2+A_2\cap CA_1,$$

de unde

$$P(A_1)=P(A_1\cap A_2)+P(A_1\cap CA_2),$$
$$P(A_2)=P(A_1\cap A_2)+P(A_2\cap CA_1),$$
$$P(A_1\cup A_2)=P(A_1\cap A_2)+P(A_1\cap CA_2)+P(A_2\cap CA_1).$$

Rezultă

$$P(A_1\cap A_2)=P(A_1)+P(A_2)-P(A_1\cup A_2). \qquad \text{(XXI.3,12)}$$

Deoarece $P(A_1)=1$, $P(A_2)=1$, deci $P(A_1\cup A_2)=1$, vom avea

$$P(A_1\cap A_2)=1.$$

Presupunînd că proprietatea este verificată pentru n mulţimi, arătăm că este verificată pentru $n+1$.

In adevăr, aplicînd (XXI.3,12) avem

$$P[(A_1\cap A_2\cap\ldots\cap A_n)\cap A_{n+1}]=P[A_1\cap A_2\ldots\cap A_n]+$$
$$+P[A_{n+1}]-P[(A_1\cap A_2\cap\ldots\cap A_n)\cup A_{n+1}],$$

de unde rezultă ceea ce trebuie arătat.

Prin aplicarea inducţiei complete începînd cu $n=2$, rezultă că propoziţia este verificată.

In cazul unei infinităţi de mulţimi, vom proceda după cum urmează.

Observăm mai întîi că

$$A_1\cap A_2\cap A_3\cap\ldots=A_1\cap(A_1\cap A_2)\cap(A_1\cap A_2\cap A_3)\cap\ldots$$

Aplicînd atunci propoziția (XXI.3,12) rezultă că

$$P\left(\bigcap_{j\in N} A_j\right)=\lim_{n\to\infty} P\left(\bigcap_{j=1}^{n} A_j\right).$$

Potrivit rezultatului precedent, $P\left(\bigcap_{j=1}^{n} A_j\right)=1$ pentru orice n, deci teorema este demonstrată în cazul general.

§ 4. Tipuri de convergență a funcției de frecvență $F_n(x)$ către $F(x)$

1. *Generalități.* Ca aplicație a teoremelor generale din capitolele precedente, vom examina aici cîteva dintre problemele mai importante relative la convergența funcției de frecvență $F_n(a)$ către funcția de distribuție $F(x)$.

Această convergență se evaluează fie ca în teoremele lui Kolmogorov și Smirnov prin $\sup\limits_{-\infty<a<\infty} |F_n(a)-F(a)|$ sau $\sup\limits_{-\infty<a<\infty} [F_n(a)-F(a)]$, fie ca în alte teoreme prin

$$\sup_{a\geqq A}\frac{|F_n(a)-F(a)|}{F(a)} \quad \text{sau} \quad \sup_{a\geqq A}\frac{F_n(a)-F(a)}{F(a)},$$

unde A este o valoare pentru care $F(A)\neq 0$, sau încă

$$\sup_{a<A'}\frac{|F_n(a)-F(a)|}{1-F(a)}, \quad \text{sau} \quad \sup_{a<A'}\frac{F_n(a)-F(a)}{1-F(a)}, \quad 1-F(A')>0.$$

De asemenea se pot evalua aceleași distanțe, însă numai pentru un interval mărginit, semnificativ pentru valorile lui $F(a)$

$$0<A_1\leqq F(a)\leqq A_2<1.$$

Vom da mai întîi cele două teoreme incipiente, în ordinea inversă a prezentării aici, utilizînd pentru demonstrarea lor metodele folosite de B. V. Gnedenko.

2. **Teorema lui Smirnov.** *Dacă $F(x)$ este continuă, crescătoare, avem*

$$\lim P\left[\sup F_n(x)<\frac{a}{\sqrt{n}}\right]=\begin{cases}1-e^{-2a^2} & a>0\\ 0 & a\leqq 0.\end{cases}$$

Demonstrație. Toată problema revine la a compara frecvențele $F_n(x)$ cu x, pentru n valori ale lui x cuprinse între 0 și 1. $y_j = F(x_j)$ fiind cele n valori, așezate în ordine crescătoare, să așezăm în ordine crescătoare valorile $x_1, x_2, \ldots, x_n$. Obținem astfel șirul

$$X_1, X_2, \ldots, X_n$$

și în același timp șirul de asemenea crescător $Y_l = F(X_l)$.
Formăm acum șirul caracteristic

$$e_1, e_2, \ldots, e_n,$$

unde punem

$$e_j = \begin{cases} 1 \text{ dacă } F_n(X_j) \geqq Y_l \\ -1 \text{ dacă } F_n(X_l) < Y_l. \end{cases}$$

Observăm că, punînd

$$s_k = e_1 + e_2 + \ldots + e_k$$

și

$$\sup [F_n(x) - F(x)] = nd_n,$$

găsim

$$\sup s_k = nd_n.$$

Dar

$$P\left(\sup_{1<k\leq n} s_k < \alpha\right) = \frac{C_n^{\left[\frac{n}{2}\right]} - C_n^{\left[\frac{n}{2}\right]-a}}{C_n^{\left[\frac{n}{2}\right]}}.$$

Dacă folosim formula lui Stirling și luăm $\alpha = a\sqrt{n}$, obținem

$$\lim P\left[\sup (F_n(a) - F(a)) < \frac{a}{\sqrt{n}}\right] = \lim P(\sup s_k < a\sqrt{n}) = 1 - e^{-2a^2},$$

pentru $a > 0$. Pentru $a \leqq 0$ teorema este evidentă.

3. Teorema lui A. N. Kolmogorov. *Dacă $F(x)$ este continuă, atunci*

$$\lim_{n\to\infty} P\left[d_n \leqq \frac{\lambda}{\sqrt{n}}\right] = D(\lambda) = \sum_{-\infty}^{+\infty} (-1)^k e^{-2k^2\lambda^2}$$

pentru $\lambda > 0$ și este nulă pentru $\lambda \leqq 0$.

Demonstrație. Să presupunem ca și înainte că variabilele $y_1, y_2, \ldots, y_n$ reprezintă rezultatele a n probe independente posibile asupra variabilei aleatorii x (ξ), așezate în ordinea crescătoare a mărimii lor.

Probabilitatea evenimentului $d_n \geqq \frac{c}{n}$, unde c este un întreg pozitiv, coincide cu aceea pentru care avem

$$F_n(x) - F(x) \geqq \frac{c}{n} \qquad \text{(XXI.4,1)}$$

sau

$$F_n(x) - F(x) \leqq -\frac{c}{n}. \qquad \text{(XXI.4,1')}$$

Dacă pentru un x este verificată inegalitatea (XXI.4,1), există un t pentru care

$$F_n(t) - F(t) = \frac{c}{n}.$$

Însă, prin definiție,

$$F_n(t) = \frac{r}{n},$$

unde r este un întreg pozitiv. Așadar, dacă pentru un anumit x are loc inegalitatea (XXI.4,1), atunci pentru variabila t corespunzătoare are loc egalitatea

$$F(t) = \frac{r-c}{n}.$$

In mod analog, dacă pentru x are loc (XXI.4,1'), atunci pentru t corespunzător avem

$$F(t) = \frac{r+c}{n}.$$

Așadar, pentru fiecare t în mod obligatoriu trebuie să existe unul dintre numerele z_k pentru care $F(z_k) = \frac{k}{n}$. Astfel, în cazul (XXI.4,1), $k = r - c$, iar în cazul (XXI.4,1'), $k = r + c$. Cu alte cuvinte, în cazul verificării inegalității (XXI.4,1) există $k+c$ valori ale șirului $y_1, y_2, \ldots, y_n$ la stînga punctului z_k

$$y_{k+c} < z_k, \quad y_{k+c+1} > z_k. \qquad \text{(XXI.4,2,}$$

In mod analog, în cazul verificării inegalității (XXI.4,1′)

$$y_{k-c} < z_k, \quad y_{k-c+1} > z_k. \tag{XXI.4,2′}$$

Evenimentul corespunzător inegalităților (XXI.4,2) se va nota cu $A_k(c)$, iar evenimentul corespunzător inegalităților (XXI.4,2′) cu $A_n(-c)$. Inegalitatea $d_n \geqq \frac{c}{n}$ este verificată atunci și numai atunci cînd se produce cel puțin unul dintre evenimentele

$$A_k(c), \quad A_k(-c) \qquad (k=1, 2, \ldots, n).$$

Să considerăm acum următoarele evenimente incompatibile

$$B_r = \left(\bigcap_{h=1, 2, \ldots, r-1} \overline{A}_h(c) \cap \overline{A}_h(-c)\right) A_r(c)$$

și

$$C_r = \left(\bigcap_{h=1, 2, \ldots, r-1} \overline{A}_h(c) \cap A_h(-c)\right) A_r(-c),$$

unde $\overline{A}$ este complementarul evenimentului A.
Deoarece

$$\bigcup_r (B_r \cup C_r) = \bigcup_k [A_k(c) \cup A_k(-c)]$$

și evenimentele B_r, C_r sînt disjuncte, avem

$$P\left\{d_n \geqq \frac{c}{n}\right\} = \sum_r [P(B_r) + P(C_r)]. \tag{XXI.4,3}$$

Pentru calculul probabilităților $P(B_r)$ și $P(C_r)$ folosim următoarele $2n$ egalități evidente

$$P(A_k(c)) = \sum_{r=1}^{k} P(B_r)\, P_{A_r(c)}(A_k(c)) + \sum_{r=1}^{k} P(C_r)\, P_{A_r(-c)}(A_k(Ac)) \tag{XXI.4,4}$$

și

$$P(A_k(-c)) = \sum_{r=1}^{k} P(B_r)\, P_{A_r(c)}(A_k(-c)) +$$

$$+ \sum_{r=1}^{k} P(C_r)\, P_{A_r(c)}(A_k(-c)).$$

Coeficienții acestor ecuații pot fi ușor calculați. Într-adevăr $P\{A_k(c)\}$ este egală cu probabilitatea ca, pentru n probe independente, variabila aleatorie să ia de $k+c$ ori o valoare mai mică decît z_k și de $n-(k+c)$ ori, o valoare mai mare decît z_k. Însă, probabilitatea variabilei aleatorii de a lua o valoare mai mică decît z_k este egală cu $\frac{k}{n}$, de aceea

$$P(A_k(c))=C_n^{k+c}\left(\frac{k}{n}\right)^{k+c}\left(1-\frac{k}{n}\right)^{n-k-c}.$$

In mod analog

$$P(A_k(-c))=C_n^{k-c}\left(\frac{k}{n}\right)^{k-c}\left(1-\frac{k}{n}\right)^{n-k+c}.$$

Probabilitatea evenimentului $A_k(c)$, cînd se cunoaște că evenimentul $A_r(c)$ a avut loc, nu este alta decît probabilitatea de apariție a $k-r$ valori de observație în intervalul (z_r, z_k) și a $(n-k-c)$ valori la dreapta lui z_k în ipoteza că $r+c$ valori se găsesc la stînga lui (z_r). Așadar,

$$P\{A_k(c)\mid A_r(c)\}=C_{n-r-c}^{k-r}\left(\frac{k-r}{n-r}\right)^{k-r}\left(1-\frac{k-r}{n-r}\right)^{n-k-c}.$$

In mod analog se obține

$$P\{A_k(c)\mid A_r(-c)\}=C_{n-r+c}^{k-r+2c}\left(\frac{k-r}{n-r}\right)^{k-r+2c}\left(1-\frac{k-r}{n-r}\right)^{n-k-c},$$

$$P\{A_k(-c)\mid A_r(c)\}=C_{n-r-c}^{k-r-2c}\left(\frac{k-r}{n-r}\right)^{k-r-2c}\left(1-\frac{k-r}{n-r}\right)^{n-k+c},$$

$$P\{A_k(-c)\mid A_r(-c)\}=C_{n-r+c}^{k-r}\left(\frac{k-r}{n-r}\right)^{k-r}\left(1-\frac{k-r}{n-r}\right)^{n-k+c}.$$

Să introducem notația

$$p_k(c)=e^{-k}\frac{k^{k+c}}{(k+c)!}. \qquad \text{(XXI.4,5)}$$

Se vede ușor că

$$P\{A_k(c)\} = \frac{p_k(c)\, p_{n-k}(-c)}{p_n(0)},$$

$$P\{A_k(-c)\} = \frac{p_k(-c)\, p_{n-k}(c)}{p_n(0)},$$

$$P\{A_k(c) \mid A_r(c)\} = \frac{p_{k-r}(0)\, p_{n-k}(-c)}{p_{n-r}(-c)},$$

$$P\{A_k(c) \mid A_r(-c)\} = \frac{p_{k-r}(2c)\, p_{n-k}(-c)}{p_{n-r}(c)},$$

$$P\{A_k(-c) \mid A_r(c)\} = \frac{p_{k-r}(-2c)\, p_{n-k}(c)}{p_{n-r}(-c)},$$

$$P\{A_k(-c) \mid A_r(c)\} = \frac{p_{k-r}(0)\, p_{n-k}(0)}{p_{n-r}(c)}.$$

În sfîrșit, mai notăm

$$u_r = P(U_r)\frac{p_n(0)}{p_{n-r}(-c)}, \qquad v_r = P(V_r)\frac{p_n(0)}{p_{n-r}(-c)}. \tag{XXI.4,6}$$

Ecuațiile (XXI.4,4) iau atunci următoarea formă

$$\begin{aligned} p_k(c) &= \sum_{r=1}^{k}\{u_r p_{k-r}(0) + v_r p_{k-r}(2c)\}, \\ p_k(-c) &= \sum_{r=1}^{k}\{u_r p_{k-r}(-2c) + v_r p_{k-r}(0)\}. \end{aligned} \tag{XXI.4,7}$$

Este important să observăm că variabilele $p_k(c)$ sînt definite pentru toți k și ecuațiile (XXI.4,7) determină pe u_r și v_r pentru toți $r>0$ (nu numai pentru $r \leqslant n$). Nu rezolvăm direct ecuațiile (XXI.4,7) ci efectuăm în prealabil transformata Laplace.

Să considerăm, pentru $\lambda>0$, funcțiile

$$u(\lambda)=\sum_{k=1}^{\infty} u_k e^{-\lambda k},$$

$$v(\lambda)=\sum_{k=1}^{\infty} v_k e^{-\lambda k},$$

$$p(\lambda,\ c)=\frac{1}{\sqrt{n}}\sum_{k=1}^{\infty} p_k(c)\, e^{-\lambda k}.$$

Ecuațiile (XXI.4,7) se reduc prin transformata Laplace la forma

$$p(\lambda,\ c)=u(\lambda)\,p(\lambda,\ 0)+v(\lambda)\,p(\lambda,\ 2c),$$

$$p(\lambda,\ -c)=u(\lambda)\,p(\lambda,\ -2c)+v(\lambda)\,p(\lambda,\ 0).$$

Să trecem acum la limită, punînd $n\to\infty$. Fie $c=z\sqrt{n}\,t>0$ arbitrar și k astfel încît pentru $n\to\infty$

$$\frac{k}{n}\to t.$$

Folosind formula lui Stirling, obținem

$$\sqrt{n}\,p_k(c)=\sqrt{n}\,e^{-k}\frac{k^{k+c}}{(k+c)!}=$$

$$=\sqrt{n}\,e^{-k}\frac{k^{k+c}}{e^{-k-c}\sqrt{2\pi\,(k+c)}\,(k+c)^{k+c}}[1+O(1)]=$$

$$=\frac{1}{\sqrt{2\pi\frac{k+c}{n}}}e^{c}\left(1+\frac{c}{k}\right)^{-c-k}(1+O(1=))\frac{1}{\sqrt{2\pi t}}e^{-\frac{z^2}{2t}}[1+O(1)],$$

deoarece $\frac{k+c}{n}\sim t$ și

$$\ln e^{c}\left(1+\frac{c}{n}\right)^{-c-k}=c-(c+k)\ln\left(1+\frac{c}{k}\right)=$$

$$=c-(c+k)\left[\frac{c}{k}-\frac{1}{2}\frac{c^2}{k^2}+O\left(\frac{c^3}{k^3}\right)\right]=$$

$$=-\frac{1}{2}\frac{c^2}{k}+O\left(\frac{c^3}{k^3}+\frac{c^4}{k^4}\right)=-\frac{1}{2}\cdot\frac{z^2}{t}+O(1).$$

In virtutea teoremei de limită pentru transformările Laplace rezultă de aici că, pentru $n\to\infty$

$$p\left(\frac{\lambda}{n}, z\sqrt{n}\right)\to\frac{1}{\sqrt{2\pi}}\int_0^\infty e^{-t\lambda-\frac{z^2}{2t}}\frac{dt}{\sqrt{t}}=\frac{1}{2\sqrt{\lambda}}e^{-z\sqrt{2\lambda}}. \qquad \text{(XXI.4,8)}$$

Deoarece

$$p\left(\frac{\lambda}{n}, -z\sqrt{n}\right)\to\frac{1}{\sqrt{2\pi}}\int_0^\infty e^{-t\lambda-\frac{z^2}{2t}}\frac{dt}{\sqrt{t}}=\frac{1}{2\sqrt{\lambda}}e^{-z\sqrt{2\lambda}}, \qquad \text{(XXI.4,8')}$$

rezultă din (XXI.4,7)

$$\lim_{n\to\infty}u\left(\frac{\lambda}{n}\right)=\lim_{n\to\infty}v\left(\frac{\lambda}{n}\right)=$$

$$=\lim_{n\to\infty}\frac{p\left(\frac{\lambda}{n}, z\sqrt{n}\right)}{p\left(\frac{\lambda}{n}, 0\right)+p\left(\frac{\lambda}{n}, 2z\sqrt{n}\right)}=\frac{e^{-z\sqrt{2\lambda}}}{1+e^{-2z\sqrt{2\lambda}}}. \qquad \text{(XXI.4,9)}$$

Pentru scopurile noastre nu avem nevoie de înseşi mărimile $P(U_k)$ şi $P(V_k)$, ci de sumele $\sum_{k=1}^{n}P(U_k)$ şi $\sum_{k=1}^{n}P(V_k)$. De aceea vom considera următoarea expresie

$$a_k=\frac{1}{p_n(0)}\sum_{r=1}^{n}p_{k-r}(-c)\,u_r \quad \text{şi} \quad b_k=\frac{1}{p_n(0)}\sum_{r=1}^{n}p_{k-r}(c)\,v_r.$$

Mărimile a_k şi b_k sînt determinate pentru toţi k (chiar pentru $k>n$). Observăm că

$$a_n=\sum_{r=1}^{n}P(U_r) \quad \text{şi} \quad b_n=\sum_{r=1}^{n}P(V_r)$$

şi deci

$$P\left\{D_n\geqq\frac{c}{n}\right\}=a_n+b_n. \qquad \text{(XXI.4,10)}$$

Să considerăm acum funcțiile (transformatele Laplace)

$$a(\lambda)=\sum_{k=1}^{\infty}a_k e^{-k\lambda}=\frac{u(\lambda)\,p(\lambda,\,-c)\,\sqrt{n}}{p_n(0)},$$

$$b(\lambda)=\sum_{k=1}^{\infty}b_k e^{-k\lambda}=\frac{v(\lambda)\,p(\lambda,\,c)\,\sqrt{n}}{p_n(0)}.$$

Luînd în considerație relația

$$\sqrt{2\pi n}\,p_n(0)\to 1 \qquad (n\to\infty)$$

și de asemenea relațiile (XXI.4,8), (XXI.4,8′), (XXI.4,9), obținem

$$\varphi(s)=\lim_{n\to\infty}\frac{1}{n}a\left(\frac{s}{n}\right)=\lim_{n\to\infty}\frac{1}{n}b\left(\frac{s}{n}\right)=$$

$$=\sqrt{\frac{\pi}{s}}\cdot\frac{e^{-2z\sqrt{2s}}}{1+e^{-2z\sqrt{2s}}}=\sqrt{\frac{\pi}{s}}\sum_{k=1}^{\infty}(-1)^{k-1}e^{-2kz\sqrt{s}}.$$

Formula (XXI.4,8) arată că funcția $\varphi(s)$ este transformata Laplace a funcției

$$f(t)=\sum_{k=1}^{\infty}(-1)^{k-1}e^{-\frac{2k^2z^2}{t}}\frac{1}{\sqrt{t}}.$$

În anumite ipoteze valabile în cazul nostru, are loc următoarea afirmație: *fie $f(t)$ o funcție nenegativă și $\varphi(s)$ transformata ei Laplace*

$$\varphi(s)=\int_0^{\infty}e^{-st}f(t)\,dt;$$

fie, mai departe, $\{u_k\}$ un șir de numere nenegative și

$$u(s)=\sum_{k=1}^{\infty}u_k e^{-ks}.$$

Atunci, dacă pentru $\delta\to 0$,

$$\delta u(s\delta)\to\varphi(s)$$

și $k\delta\to t$, avem pentru orice $t>0$

$$u_k\to f(t).$$

In cazul nostru $\delta = \frac{1}{n}$, de aceea trebuie să avem

$$\lim_{n\to\infty} a_n = \lim_{n\to\infty} b_n = f(1).$$

Aşadar,

$$\lim_{n\to\infty}(a_n+b_n)=2\sum_{k=1}^{\infty}(-1)^{k-1}e^{-2k^2z^2}=1-\sum_{k=-\infty}^{\infty}(-1)^k e^{-2k^2z^2}.$$

Luînd în considerație egalitatea (XXI.4,10) obținem în sfîrșit

$$\lim_{n\to\infty} P\{\sqrt{n}D_n<z\}=\sum_{k=-\infty}^{\infty}(-1)^k e^{-2k^2z^2}=K(z).$$

4. Alte teoreme. Vom demonstra mai întîi că, *pentru* $a>0$ *şi* $0<\alpha<1$,

$$\lim P\left(\sup_{\alpha\le F_n(x)}\frac{F_n(x)-F(x)}{F(x)}<\frac{a}{\sqrt{n}}\right)=\sqrt{\frac{2}{\pi}}\int_6^{a\sqrt{\frac{\alpha}{1-\alpha}}}e^{-\frac{u^2}{2}}\,du. \quad \text{(XXI.4,11)}$$

Este evident că

$$\sup_{\alpha\le F_n(x)}\frac{F_n(x)}{F(x)}=\sup_{\alpha n\le k\le n}\frac{\frac{k}{n}}{Y_k}.$$

Trebuie dar să evaluăm

$$P\left(\sup_{\alpha n\le k\le n}\frac{\frac{k}{n}-Y_k}{Y_k}<\frac{a}{\sqrt{n}}\right).$$

Vom folosi, pentru aceasta, teorema relativă la suma de variabile independente. De acolo rezultă că

$$\lim P\left\{\sup_{\alpha n\le k\le n}\left(z_k-\sum_{m=k}^{n}\frac{1}{m}\right)<\xi\sqrt{\sum\frac{1}{k^2}}\right\}=\sqrt{\frac{2}{\pi}}\int_0^{\xi}e^{-\frac{u^2}{2}}du.$$

Deoarece

$$\sum_{m=k}^{n}\frac{1}{m}=\ln\frac{n}{k}+O\left(\frac{1}{n}\right)$$

și

$$\sqrt{\sum_{\alpha n \leq k \leq n} \frac{1}{k^2}} = \sqrt{\frac{1-\alpha}{n\alpha}} + O\left(\frac{1}{n}\right),$$

rezultă de mai sus că

$$\lim P\left\{\sup_{\alpha n \leq k \leq n} \ln\left(\frac{\frac{k}{n}}{Y_k}\right) < \xi\sqrt{\frac{1-\alpha}{n\alpha}}\right\} =$$

$$= \lim P\left\{\sup_{\alpha n \leq k \leq n} \ln\left(\frac{\frac{k}{n}}{Y_k} - 1\right) < \frac{a}{\sqrt{n}}\right\} = \sqrt{\frac{2}{\pi}} \int_0^{a\sqrt{\frac{\alpha}{1-\alpha}}} e^{-\frac{u^2}{2}}\, du,$$

deci formula (XXI.4.11).

Din aceasta rezultă în primul rînd următoarea

Teoremă.

$$\lim P\left\{\sup_{F(x) \geq \alpha} \frac{F_n(x) - F(x)}{F(x)} < \frac{a}{\sqrt{n}}\right\} = \sqrt{\frac{2}{\pi}} \int_0^{a\sqrt{\frac{\alpha}{1-\alpha}}} e^{-\frac{u^2}{2}}\, du \quad \text{(XXI.4,12)}$$

pentru $a > 0$. *Pentru* $a < 0$, *limita menbrului întîi este nulă.*

In adevăr, introducînd următoarele mulțimi notate pentru simplificarea scrisului cu cîte o literă

$$A^{(n)} = \left(\sup_{F(x) \geq \alpha} \frac{F_n(x) - F(x)}{F(x)} < \frac{a}{\sqrt{n}}\right),$$

$$A_1^{(n)} = \left(\sup_{F_n(x) \geq \alpha + \varepsilon} \frac{F_n(x) - F(x)}{F(x)} < \frac{a}{\sqrt{n}}\right),$$

$$A_2^{(n)} = \left(\sup_{F_n(x) \geq \alpha - \varepsilon} \frac{F_n(x) - F(x)}{F(x)} < \frac{a}{\sqrt{n}}\right),$$

$$B^{(n)} = (|F_n(x) - F(x)| \leq \varepsilon),$$

deoarece

$$A^{(n)} \cap B^{(n)} \subset A_1^{(n)} \cap B^{(n)},$$

$$A_2^{(n)} \cap B^{(n)} \subset A^{(n)} \cap B^{(n)},$$

avem

$$P(A^{(n)}) = P(A^{(n)} \cap B^{(n)}) + P(A^{(n)} \cap CB^{(n)}) \leq P(A_1^{(n)} \cap B^{(n)}) + P(CB^{(n)}) \leq P(A_1^{(n)}) + P(CB^{(n)})$$

şi de asemenea

$$P(A_2^{(n)}) \leq P(A^{(n)}) + P(CB^{(n)}),$$

adică

$$P(A_2^{(n)}) - P(CB^{(n)}) \leq P(A^{(n)}) \leq P(A_1^{(n)}) + P(CB^{(n)}).$$

Intrucît, după teorema lui Glivenko,

$$\lim_{n\to\infty} P(CB^{(n)}) = 0,$$

iar egalitatea (XXI.4,11) arată că $\lim_{n\to\infty} P(A_1^{(n)})$ şi $\lim_{n\to\infty} P(A_2^{(n)})$ există şi sînt date de integralele din membrul al doilea din (XXI.4,11) înlocuind pe α respectiv cu $\alpha+\varepsilon$ şi $\alpha-\varepsilon$; cum ε este arbitrar de mic şi integrala este continuă de limita superioară, vom avea

$$\lim_{n\to\infty} P(A^{(n)}) = \sqrt{\frac{2}{\pi}} \int_0^{a\sqrt{\frac{\alpha}{1-\alpha}}} e^{-\frac{u^2}{2}}\, du,$$

deci egalitatea enunţată în teoremă.

Vom demonstra, în sfîrşit, şi următoarea

Teoremă.

$$\lim P\left(\sup_{F(x)\geq\alpha} \frac{|F_n(x)-F(x)|}{F(x)} < \frac{a}{\sqrt{n}}\right) =$$

$$= \frac{4}{\pi} \sum_{l=0}^{\infty} \frac{(-1)^l}{2l+1} e^{-\frac{(2k+1)^2\pi^2}{8}\frac{1}{\alpha a^2}}$$

pentru $a>0$. *Pentru* $a<0$, *probabilitatea limită este nulă.*

Să observăm, pentru aceasta, că

$$\sup_{F_n(x)\geq\alpha} \frac{|F_n(x)-F(x)|}{F(x)} = \sup_{n\alpha\leq k\leq n} \left(\frac{\left|\frac{k}{n} - Y_k\right|}{Y_k}, \frac{\left|\frac{k}{n} - Y_{k-1}\right|}{Y_k} \right).$$

De aici urmează

$$\sup\left|\frac{k}{nY_k}-1\right| \leqslant \sup\left|\frac{F_n(x)-F(x)}{F(x)}\right| \leqslant$$

$$\leqslant \sup\left|\frac{k}{nY_k}-1\right|+\frac{1}{\alpha n}, \qquad k \leqslant \alpha n.$$

Dar, în aceleaşi condiţii pentru k, avem

$$\left|\frac{k}{nY_{k+1}}-1\right| \leqslant \left|\frac{k+1}{nY_{k+1}}-1\right|+\frac{1}{\alpha n}$$

şi întrucît

$$\sup\left|\frac{k+1}{nY_{k+1}}-1\right| \leqslant \sup\left|\frac{k}{nY_k}-1\right|,$$

rezultă

$$\sup\left(\left|\frac{k}{nY_k}-1\right|, \left|\frac{k}{nY_{k+1}}-1\right|\right) \leqslant \sup\left|\frac{k}{nY_k}-1\right|+\frac{1}{\alpha n}.$$

Avem prin urmare de evaluat distribuţia limită a variabilei

$$\sqrt{n}\sup\left|\frac{k}{nY_k}-1\right|.$$

In acelaşi mod ca în cazul teoremei precedente, se arată că distribuţia limită a acestei variabile este identică cu aceea a variabilei

$$\sqrt{n}\sup_{\alpha n \leqslant t \leqslant n}\left|\ln\frac{1}{Y_k}\sum_{m=k}^{n}\frac{1}{m}\right| = \sqrt{n}\sup_{\alpha n \leqslant k \leqslant n}\left|\sum_{k=1}^{n-k+1}\frac{\theta_h-1}{n+1-h}\right|$$

întrebuinţînd notaţiile paragrafului precedent.

Aplicînd acum egalitatea (XXI.4,11) din teorema 2 din paragraful precedent, obţinem rezultatul enunţat în teoremă.

§ 5. Compararea a două funcţii de frecvenţă

Dacă $x'_1, x'_2, \ldots, x'_n; x''_1, x''_2, \ldots, x''_n$ sînt două şiruri de cîte n valori ale variabilei x, a cărei funcţie de distribuţie continuă crescătoare este $F(x)$, notăm cu $F'_n(x)$ şi $F''_n(x)$ cele două funcţii de frecvenţă corespunzătoare celor două şiruri.

O *teoremă a lui Smirnov* are următorul enunț

$$\lim_{n\to\infty} P\left\{\sup_{-\infty<x<\infty}(F_n'(x)-F_n''(x))<\frac{a}{\sqrt{\frac{n}{2}}}\right\}=\begin{cases}1-e^{-2a^2} & \text{pentru } a>0\\ 0 & \text{pentru } a<0,\end{cases} \tag{XXI.5,1}$$

$$\lim_{n\to\infty} P\left\{\sup_{-\infty<x<\infty}|F_n'(x)-F_n''(x)|<\frac{a}{\sqrt{\frac{n}{2}}}\right\}=$$

$$=\begin{cases}\sum\limits_{k=-\infty}^{\infty}(-1)^k e^{-2k^2a^2} & \text{pentru } a>0\\ 0 & \text{pentru } a<0\end{cases} \tag{XXI.5,2}$$

ca în cazul comparării unei singure funcții de frecvență cu $F(x)$. In vederea stabilirii acestei teoreme vom pune

$$d_n=\sup_{-\infty<x<\infty}(F_n'(x)-F_n''(x)),$$

$$\delta_n=\sup_{-\infty<x<\infty}|F_n'(x)-F_n''(x)|$$

și vom începe prin a demonstra următoarea *teoremă a lui Gnedenko*

$$P\left\{d_n<\sqrt{\frac{2}{n}}\,a\right\}=1-\frac{C_{2n}^{n-a\sqrt{2n}}}{C_{2n}^{n}},$$

$$P\left\{\delta_n<\sqrt{\frac{2}{n}}\,a\right\}=\frac{1}{C_{2n}^{n}}\sum_{k=-\left[\frac{n}{2}\right]}^{\left[\frac{n}{2}\right]}(-1)^k C_{2n}^{n-ka\sqrt{2n}}, \tag{XXI.5,3}$$

pentru $\frac{1}{\sqrt{2n}}<a\leq\sqrt{\frac{n}{2}}$ și

$$P\left\{d_n<\sqrt{\frac{2}{n}}\,a\right\}=P\left\{\delta_n<\sqrt{\frac{2}{n}}\,a\right\}=\begin{cases}0, & z\leq\frac{1}{\sqrt{2n}}\\ 1, & z>\frac{1}{\sqrt{2n}}.\end{cases} \tag{XXI.5,4}$$

Fie $X_1, X_2, \ldots, X_{2n}$ șirul valorilor x_i', x_j'' $(i=1, 2, \ldots, n;$ $j=1, 2, \ldots, n)$ ordonat crescător.

Constituim un șir caracteristic $e_1, e_2, \ldots, e_{2n}$ al celor două sisteme date în modul următor:
e_j este egal cu $+1$ sau cu -1, după cum X_j aparține primului sau celui de al doilea șir.
Dacă

$$s_k = e_1 + e_2 + \ldots + e_k,$$

avem, evident,

$$\sup_{1 \leq k \leq 2n} s_k = nd_n, \qquad \sup_{1 \leq k \leq 2n} |s_k| = n\delta_n.$$

Nu avem acum decît să evaluăm probabilitățile

$$P(nd_n < \alpha) = P\left(\sup_{1 \leq k \leq 2n} s_k < \alpha\right),$$

$$P(n\delta_n < \alpha) = P\left(\sup_{1 \leq k \leq 2n} |s_k| < \alpha\right)$$

cu $\alpha \leq n$.

Pentru calculul acestor probabilități ne folosim de imaginea geometrică a mersului la întîmplare pe un carelaj, cuprinzînd pe axa x punctele $(0, 0)$; $(\pm 1, 0), \ldots, (\pm n, 0)$, iar pe latura paralelă la distanța $2n$ punctele $(0, 2n)$, $(\pm 1, 2n), \ldots, (\pm n, 2n)$.
Observăm că $s_{2n} = 0$, deoarece numărul pașilor pozitivi și negativi este același. Prin urmare, fiecare drum de considerat pleacă din $(0, 0)$ și termină în $(0, 2n)$. Numărul tuturor drumurilor distincte este evident C_{2n}^{n}.
Pe de altă parte, numărul tuturor drumurilor care depășesc ordonata α este evident $C_{2n}^{n-\alpha} = C_{2n}^{n+\alpha}$. Deci numărul drumurilor care nu depășesc ordonata α este $C_{2n}^{n} - C_{2n}^{n-\alpha}$. De unde, prima parte a teoremei

$$P(nd_n < \alpha) = P\left(\sup_{1 \leq k \leq 2n} s_k < \alpha\right) = 1 - \frac{C_{2n}^{n-\alpha}}{C_{2n}^{n}},$$

pentru $0 < \alpha \leq n$.

Evaluînd numărul de drumuri care pleacă din $(0, 0)$, ajung în $(0, 2n)$ și nu depășesc la dreapta verticala în $(\alpha, 0)$ iar la stînga verticala în $(-\alpha, 0)$, obținem

$$P(n\delta_n<\alpha)=P\Big(\sup_{1\leq k\leq 2n}|s_k|<\alpha\Big)=1+\sum_{k=1}^{\left[\frac{n}{2}\right]}(-1)^k\frac{C_{2n}^{n-k\alpha}}{C_{2n}^{n}},$$

de asemenea pentru $0<\alpha\leq n$.

Este acum ușor să verificăm teoremele limită ale lui Smirnov.

Punem $\alpha=a\sqrt{2n}$. Expresia

$$A_{\alpha,n}=\frac{C_{2n}^{n-\alpha}}{C_{2n}^{n}}=\frac{(n!)^2}{(n-\alpha)!\,(n+\alpha)!}$$

dă, prin aplicarea formulei lui Stirling,

$$A_{\alpha,n}=\left(1-\frac{\alpha}{n}\right)^{-n+\alpha}\left(1+\frac{\alpha}{n}\right)^{-n-\alpha}(1+o(1)).$$

De unde

$$\ln A_{\alpha,n}=-\frac{\alpha^2}{n}+o(1)=-2\alpha^2+o(1).$$

Iar dacă $\alpha\to ak\sqrt{2n}$, obținem

$$\ln A_{ka,n}=-2k^2a^2+o(1)$$

și cu aceasta ambele formule sînt verificate. Este numai necesar să se observe că suma din membrul al doilea al ultimei formule se transformă la limită într-o serie absolut și uniform convergentă.

Demonstrarea părții finale a teoremei se lasă în seama cititorului.

Observație finală. Teorema lui Gnedenko este foarte bogată în conținut, întrucît ne dă expresia exactă a legii chiar pentru n finit. Este deci mai susceptibilă de a fi comparată cu datele experimentale care sînt de cele mai multe ori puțin numeroase sau, în orice caz, nu așa de mari ca să le putem compara cu o lege asimptotică.

Vom compara două șiruri cu același număr de valori. Fie

$$x_1, x_2, \ldots, x_n; \quad y_1, y_2, \ldots, y_n$$

două șiruri de cîte n valori ale variabilei x.

Să aranjăm aceste șiruri în ordinea mărimii

$$z_1 < z_2 < \ldots < z_{2n} \qquad \text{(XXI.5,5)}$$

și să introducem variabilele aleatorii

$$\varepsilon_r = \begin{cases} \dfrac{1}{\sqrt{2n}} & \text{dacă} \quad z_r = y_i \\ -\dfrac{1}{\sqrt{2n}} & \text{dacă} \quad z_r = x_j. \end{cases} \qquad \text{(a)}$$

Să considerăm sumele

$$S_k = \sum_{r=1}^{k} \varepsilon_r \qquad (k = 1, 2, \ldots, N), \quad S_0 = 0. \qquad \text{(b)}$$

Se vede ușor că

$$S_k = \sqrt{\frac{n}{2}} \left(F'_n(t) - F''_n(t)\right), \qquad \text{(c)}$$

iar t se definește ca cea mai mică rădăcină a ecuației $K_N(t) = k$, unde $K_N(t)$ este numărul de observații din șirul (XXI.5,5), mai mici decît t.

Să introducem distribuțiile condiționate

$$p_n(x, t, z) = P_{S_k = x}\left\{\max_{r \leqq k} S_r < z\right\}, \qquad \text{(XXI.5.6)}$$

$$q_n(x, t, z) = P_{S_k = x}\left\{\max_{r \leqq k} |S_r| < z\right\}, \qquad \text{(XXI.5,7)}$$

unde t ia valori de forma $\frac{k}{N}$ $(k = 1, 2, .., N)$, iar x toate valorile posibile ale sumelor S_k. Funcțiile $q_n(x, t; a)$ și $p_n(x, t; a)$ sînt definite într-un număr finit de puncte ale planului (x, t). În schema de mers la întîmplare corespunzătoare sumelor S_k în planul (x, t), punctele în care sînt definite funcțiile $p(x, t; a)$ și $q(x, t; a)$ sînt punctele nodale ale drumurilor. Coordonatele lor au forma

$$x = r\alpha - s\beta, \quad t = \frac{k}{2n}, \quad k = r + s, \quad r, s \geqq 0. \qquad \text{(XXI.5,8)}$$

Din definiția funcțiilor $p(x, t; a)$ și $q(x, t; a)$ rezultă că

$$\Phi_n(a)=P\left\{\sqrt{\frac{n}{2}}\ \sup\ (F_n'(t)-F_n''(t))<a\right\}=p(0,\ 1,\ a),$$
$$\Psi_n(a)=P\left\{\sqrt{\frac{n}{2}}\ \sup\ |F_n'(t)-F_n''(t)|<a\right\}=q(0,\ 1,\ a), \qquad \text{(XXI.5,9)}$$

Pentru determinarea funcțiilor $p(x, t; a)$ și $q(x, t; a)$ să considerăm schema mersului la întîmplare, determinată de sumele S_k. In această schemă, S_k determină abscisa punctului care se mișcă în planul (x, t) și care face la fiecare pas un salt pe axa t în valoare egală cu $\frac{1}{2p}$ și pe axa x un salt egal cu $\frac{1}{\sqrt{2n}}$ sau $-\frac{1}{\sqrt{2n}}$. Deoarece $S_0=0$, $S_N=0$, toate drumurile pornesc din punctul $(0, 0)$ și sfîrșesc în punctul $(0, 1)$. Se înțelege ușor că numărul tuturor drumurilor posibile este egal cu $C_{2n}^n=\frac{(2n)!}{(n!)^2}$ și toate drumurile sînt egal probabile. Dacă notăm cu $u^+(x, t; a)$ numărul drumurilor care ajung în punctul (x, t) și nu au puncte comune cu frontiera $x=z$, obținem

$$p(x, t; a)=P_{S_k=x}\{\max S_r<a\}. \qquad \text{(d)}$$

Insă

$$P_{S_k=x}\left\{\max_{r<k} S_r<a\right\}=\frac{u(x, t, a)\,C_{2n-k}^{n-r}}{C_{2n}^n}, \qquad \text{(e)}$$

unde r este numărul de pași pozitivi pînă la punctul (x, t), iar

$$P\{S_k=x\}=\frac{C_k^r C_{2n-k}^{n-r}}{C_{2n}^n}. \qquad \text{(f)}$$

Așadar obținem

$$p_n(x, t, a)=\frac{u(x, t, a)}{C_k^r} \qquad \text{(XXI.5,10)}$$

și în mod analog

$$q_n(x, t, a)=\frac{v(x, t, a)}{C_k^r}, \qquad \text{(XXI.5.11)}$$

unde $u(x, t, z)$ este numărul de drumuri care vin în punctul (x, t) și nu au puncte comune cu frontierele $x=z$ și $x=-z$.

Funcțiile $u(x, t; a)$ și $v(x, t; a)$ satisfac următoarea ecuație cu diferențe

$$u(x, t, a)=u\left(x-\frac{1}{\sqrt{2n}},\ t-\frac{1}{2n},\ a\right)+u\left(x+\frac{1}{\sqrt{2n}},\ t-\frac{1}{2n},\ a\right), \quad \text{(g)}$$

care este valabilă pentru punctele interioare domeniului Q, mărginit de dreptele

$$t=-\frac{1}{\sqrt{2n}}x, \quad t=\frac{1}{\sqrt{2n}}x, \quad t=\frac{1}{\sqrt{2n}}x+1, \quad t=-\frac{1}{\sqrt{2n}}x+1. \quad \text{(h)}$$

Din (XXI.5,10) (XXI.5,11) și (h) rezultă că funcțiile $p(x, t; z)$ și $q(x, t; z)$ satisfac în interiorul domeniului Q, ecuația cu diferențe

$$\pi(x, t)=\frac{r}{k}\pi\left(x-\frac{1}{\sqrt{2n}},\ t-\frac{1}{2n}\right)+\left(1-\frac{r}{k}\right)\pi\left(x+\frac{1}{\sqrt{2n}},\ t-\frac{1}{2n}\right), \quad \text{(i)}$$

unde am suprimat, pentru simplitate, dependența de a.

Deoarece numerele x, t, k și r sînt egale prin ecuațiile (XXI.5,8), avem

$$r=nt+\gamma x, \quad k=2nt, \quad \text{(j)}$$

unde

$$\gamma=\sqrt{\frac{n}{2}}. \quad \text{(k)}$$

Ecuația cu diferențe (XXI.5,11) se mai poate scrie și sub forma

$$\Omega(\pi)=-\pi(x, t)+\frac{1}{2}\left(1+\frac{1}{\sqrt{2n}}\cdot\frac{x}{t}\right)\pi\left(x-\frac{1}{\sqrt{2n}},\ t-\frac{1}{2n}\right)+$$

$$+\frac{1}{2}\left(1-\frac{1}{\sqrt{2n}}\cdot\frac{x}{t}\right)\pi\left(x+\frac{1}{\sqrt{2n}},\ t-\frac{1}{2n}\right)=0. \quad \text{(XXI.5,12)}$$

Din definiția funcției $p_n(x, t; z)$, rezultă că

$$\begin{aligned} p_n\left(x,\ \frac{1}{\sqrt{2n}}x,\ a\right)&=1 \quad \text{pentru} \quad 0\leqq x<a,\\ p_n\left(x,\ -\frac{1}{\sqrt{2n}},\ a\right)&=1 \quad \text{pentru} \quad x<0, \quad \text{(XXI.5,13)}\\ p_n(x, t, a)&=0 \quad \text{pentru} \quad x\geqq a. \end{aligned}$$

Deci funcția $p(x, t; z)$ este soluția ecuației (XXI.5,12) cu condițiile limită (XXI.5,13).

In mod analog, rezultă că funcţia $q_n(x, t; z)$ este soluţia ecuaţiei (XXI.5,12) cu condiţiile limită

$$\begin{aligned} q_n\left(x, \quad \frac{1}{\sqrt{2n}}x, \ a\right)&=1 \quad \text{pentru} \quad 0\leqq x<a, \\ q_n\left(x, \ -\frac{1}{\sqrt{2n}}x, \ a\right)&=1 \quad \text{pentru} \quad -z<x<0, \\ q_n(x, t, z) \quad &=0 \quad \text{pentru} \quad |x|\geqq a. \end{aligned} \tag{XXI.5,14}$$

Dezvoltările asimptotice ale funcţiilor $p_n(x, t; a)$, $q_n(x, t; a)$ dau dezvoltările asimptotice ale funcţiilor $\Phi_n(a)$ şi $\Psi_n(a)$.

Reprezentările asimptotice ale funcţiilor $p_n(x, t; a)$ şi $q_n(x, t; a)$ se găsesc căutînd soluţia ecuaţiei (XXI.5,12) sub forma

$$\pi(x, t)=\sum_{l=0}^{\infty}\left(\frac{1}{\sqrt{2n}}\right)^l \pi_l(x, t). \tag{XXI.5,15}$$

După formula lui Taylor avem

$$\pi\left(x-\varepsilon\frac{1}{\sqrt{2n}}, \ t-\frac{1}{2n}\right)=\pi(x, t)+$$

$$+\sum_{k=1}^{\infty}\frac{(-1)^k}{(\sqrt{2n})^k k!}\left(\frac{\partial}{\partial x}+\varepsilon\frac{1}{\sqrt{2n}}\frac{\partial}{\partial t}\right)^{(k)}\pi(x, t) \tag{l}$$

$$\varepsilon=\pm 1.$$

Inlocuind expresiile obţinute în ecuaţia (XXI.5,12), obţinem

$$\Omega(\pi)=\frac{1}{2}\left(1+\frac{1}{\sqrt{2n}}\cdot\frac{x}{t}\right)\sum\frac{(-1)^k}{(\sqrt{2n})^k \cdot k!}\left(\frac{\partial}{\partial x}+\frac{1}{\sqrt{2n}}\cdot\frac{\partial}{\partial t}\right)^{(k)}\pi(x, t)+$$

$$+\frac{1}{2}\left(1-\frac{1}{\sqrt{2n}}\frac{x}{t}\right)\sum\frac{(-1)^k}{(\sqrt{2n})^k k!}\left(\frac{\partial}{\partial x}+\frac{1}{\sqrt{2n}}\frac{\partial}{\partial t}\right)^{(k)}\pi(x, t)=0. \tag{n}$$

Dacă, acum, folosim (XXI.5,15) găsim $\pi_0(x, t)$, izolînd termenii care nu cuprind pe $\frac{1}{\sqrt{2n}}$; ecuaţia obţinută este

$$\frac{\partial \pi_0}{\partial t} = -\frac{x}{t}\frac{\partial \pi_0}{\partial x} + \frac{1}{2}\frac{\partial^2 \pi_0}{\partial x^2},$$

cu condiţiile iniţiale

$$\pi_0(x, 0)=1, \quad \pi_0(a, t)=0.$$

Soluţia acestei ecuaţii de difuziune binecunoscută este

$$\pi_0(x, t, a)=1-e^{-2\frac{u(a-x)^2}{t}}. \qquad \text{(XXI.5,16)}$$

Capitolul XXII

Statistica ordinei

§ 1. Statistica de ordine pentru variabile exponenţiale şi omogene

1. Obiectul acestui capitol al Statisticii matematice este studiul sistemelor de valori ale unei variabile aleatorii $x=f(\xi)$ din punctul de vedere al relaţiei lor de ordine.

Aşa, de pildă, studiul celei mai mari sau al celei mai mici dintre aceste valori, sau studiul *extinderii (range)* care separă pe cea mai mare de cea mai mică dintre aceste valori, considerate ca variabile aleatorii, sînt probleme ale statisticii ordinei. In expunere s-a urmat mai cu seamă modul de prezentare al lui Rényi.

2. Justificarea acestei Statistici. Acest capitol special se sprijină pe două observaţii foarte simple.

Dacă funcţia de distribuţie $F(x)$ a variabilei aleatorii $x=x(\xi)$ este continuă, funcţia de distribuţie a variabilei $F(x(\xi))$ are forma unică

$$\Psi(a)=\begin{cases} 0, & a<0, \\ a, & 0\leq a\leq 1, \\ 1, & 1<a, \end{cases}$$

Demonstrația este imediată:

Funcția continuă $F(x)$ este măsurabilă, deci $F(x(\xi))$ este o variabilă aleatorie.

Este, pe de altă parte, evident că mulțimea evenimentelor elementare ξ pentru care $y=F(x(\xi))$ este mai mic de cît zero este vidă, deci probabilitatea ei este nulă.

De asemenea, este evident că toate evenimentele ne dau $F(x)\leqslant 1$. Deci probabilitatea de a avea $F(x)\leqslant 1$ este egală cu 1.

Pe de altă parte, dacă $0\leqslant a\leqslant 1$ și presupunem că $F(\alpha)=a$, α fiind cea mai mică valoare pentru care avem această egalitate,

$$\Psi(a)=P\{\xi\,|\,F(x(\xi))<a\}=P\{\xi\,|\,x(\xi)<\alpha\}=F(\alpha)=a.$$

Dacă funcția de distribuție $F(x)$ este continuu crescătoare, variabila aleatorie $z=\ln\frac{1}{F(x(\xi))}$ are o distribuție exponențială.

Observația rezultă imediat din precedenta. Se poate stabili și direct. $z=z(\xi)$ este evident o variabilă aleatorie odată cu $x(\xi)$.

Avem, evident, pentru $a\geqslant 0$

$$P(z>a)=P\left(\ln\frac{1}{F(x)}>a\right)=P\left(\frac{1}{F(x)}>e^{a}\right)=$$

$$=P(F(x)<e^{-a})=P(x<F^{-1}(e^{-a}))=F(F^{-1}(e^{-a}))=e^{-a},$$

de unde

$$P(z<a)=1-e^{-a}\qquad(a\geqslant 0).$$

Deci, notînd cu $\chi(a)$ funcția de distribuție a variabilei z, avem

$$\chi(a)=\begin{cases}1-e^{-a} & a\geqslant 0\\ 0 & a<0.\end{cases}$$

3. Fie $x_1, x_2, \ldots, x_n$ un sistem de valori independente ale variabilei aleatorii $x=x(\xi)$, a cărei funcție de distribuție $F(x)$ este continuu crescătoare.

Introducem concomitent cele două variabile $y=F(x(\xi))$ și $z=\ln\frac{1}{F(x(\xi))}$ deci, odată cu $x_1, x_2, \ldots, x_n$, valorile

$$y_j=F(x_j),\quad z_j=\ln\frac{1}{y_j}\qquad(j=1, 2, \ldots, n).$$

Potrivit observaţiilor 1 şi 2, variabila y are funcţia de distribuţie uniformă $\psi(a)$, iar variabila z, funcţia exponenţială $\chi(a)$.
Dacă notăm cu $X_1, X_2, \ldots, X_n$ sistemul $x_1, x_2, \ldots, x_n$ ordonat crescător $X_1 \leqq X_2 \leqq \ldots \leqq X_r$, avem concomitent sistemele ordonate crescător

$$Y_k = F(X_k), \quad z_k = \ln \frac{1}{Y_{n-k+1}} \qquad (k=1, 2, \ldots, n). \quad \text{(XXII.1,1)}$$

4. In acest paragraf ne vom ocupa de sistemele $z_1, z_2, \ldots, z_n$, cu distribuţie exponenţială şi vom stabili legi de distribuţie asimptotică pentru variabilele sistemului ordonat crescător $z_1, z_2, \ldots, z_n$, precum şi de sistemele $y_1, y_2, \ldots, y_n$, $Y_1, Y_2, \ldots, Y_n$ respectiv.

Teorema 1. *Succesiunea $z_1, z_2, \ldots, z_n$ constituie un lanţ cu legături complete cu creşteri independente.*

Funcţia de distribuţie a diferenţei $z_{l+1} - z_l$ satisface, evident, relaţia funcţională

$$\underset{z_l \geqq b}{P}(z_{l+1} - z_l > a) = \underset{z_l \geqq b}{P}(z_{l+1} > a+b).$$

Să observăm că z_l şi z_k fiind două variabile cu distribuţie exponenţială, $\Phi(a) = 1 - e^{-a}$, avem

$$P(z_h < x+y) = \frac{P(z_h < x+y,\ z_k < y)}{P(z_k < y)} = \frac{\Phi(x+y) - \Phi(y)}{1 - \Phi(y)} = 1 - e^{-x}.$$

Probabilitatea este aceeaşi pentru toate valorile lui h pentru care $z_h > z_k$; deoarece aceste variabile sînt independente, probabilitatea de a avea $z_{k+1} > x+y$, $z_{k+2} > x+y, \ldots, z_n > x+y$, dacă $z_k > y$, este $e^{-(n-k)x}$; avem deci

$$\underset{z_l > y}{P}(z_{l+1} \geqq x+y) = e^{-(n-l)x}, \qquad \text{(XXII.1,2)}$$

presupunînd că $z_l = z_k$, deoarece aceasta indicînd ordinea crescătoare ne arată implicit că este verificată şi pentru $z_{l+2}, \ldots, z_n$.
Din (XXII.1,2) rezultă prin urmare

$$\underset{z_l \geqq b}{P}(z_{l+1} - z_l < a) = 1 - e^{-(n-l)a}. \qquad \text{(XXII.1,3)}$$

Dar atunci, întrucît momentul de ordinul al doilea nu depinde de b, rezultă că

$$P(z_{l+1} - z_l < a) = 1 - e^{-(n-l)a}. \qquad \text{(XXII.1,4)}$$

Deci, diferența $z_{l+1}-z_l$ are o distribuție exponențială cu valoarea medie $\frac{1}{n-l}$.

Variabilele

$$\theta_{l+1}=(n-l)(z_{l+1}-z_l) \qquad \text{(XXII.1,5)}$$

sînt independente între ele. In adevăr, din relațiile (XXII.1,2), (XXII.1,3), (XXII.1,4) rezultă

$$P_{z_1=b_1,\ z_2-z_1=b_2,\ \ldots,\ z_l-z_{l-1}=b_l}(z_{l+1}-z_l<a)=$$

$$=P_{z_1=b_1,\ z_2=b_1+b_2,\ \ldots.\ z_l=b_1+b_2+\ldots+b_l}(z_{l+1}-z_l<a)=$$

$$=P(z_{l+1}-z_l<a)=1-e^{-(n-l)a}.$$

Deci variabilele $z_{l+1}-z_l$, de unde variabilele θ_l, sînt independente între ele și au legi de distribuție exponențiale. θ_l are o lege de distribuție exponențială cu valoarea medie 1.

Din (XXII.1,4) rezultă

$$z_l=\frac{\theta_1}{n}+\frac{\theta_2}{n-1}+\ldots+\frac{\theta_l}{n+l+1}, \qquad \text{(XXII.1,6)}$$

expresie care arată, că în adevăr, variabilele z_l constituie un lanț cu legături complete cu creșteri independente.

Teorema 2. *Variabila nz_l are, pentru l fix și $n\to\infty$, o distribuție Γ de ordinul l*

$$\lim_{n\to\infty}P(nz_l<a)=\frac{1}{(l-1)!}\int_0^a u^{l-1}e^{-u}\,du. \qquad \text{(XXII.1,7)}$$

Pentru demonstrație, să observăm mai întîi că variabila z_l reprezentată în (XXII.1,6) este o sumă de variabile independente $\frac{\theta_h}{n-h+1}$ cu densitățile respective $(n-h+1)e^{-(n-h+1)a}$ și cu funcțiile caracteristice

$$f_{h,n}(\tau)=(n-h+1)\int_0^\infty e^{i\tau a}e^{-(n-h+1)a}\,da.$$

z_l are deci funcţia caracteristică

$$g_{l,n}(\tau)=\prod_{h=1,\,2,\,\ldots,\,l} f_{h,l}(\tau)=$$

$$=n(n-1)\ldots(n-l+1)\int_0^\infty\ldots\int_0^\infty e^{i\tau(a_1+a_2+\ldots+a_l)}\,e^{-\sum_{h=1}^{l}(n-h+1)a_h}\,da_1\ldots da_l=$$

$$=l!\,C_n^l\int_0^\infty\ldots\int_0^\infty e^{i\tau A}\,e^{-(n-l+1)A}\,e^{-(l-1)A}\,e^{-(l-2)A}\ldots e^{-a_l-1}da_1\ldots da_l,$$

unde $A=a_1+a_2+\ldots+a_l$. Făcînd schimbarea de variabilă

$$A_1=a_1,\quad A_2=a_2,\ldots,A_{l-1}=a_{l-1},\quad A=a_1+a_2+\ldots+a_l,$$

şi observînd că

$$\int_0^\infty e^{-(l-h)A_h}\,dA_h=\frac{1}{l-h}\qquad(h=1,\,2,\ldots,l-1),$$

găsim

$$g_{l,n}(\tau)=lC_n^l\int_0^\infty e^{i\tau A}\,e^{-(n-l+1)A}\,(1-e^{-A})^{l-1}\,dA.$$

Funcţia de distribuţie a variabilei nz_l va fi

$$g_{l,n}(n\tau)=lC_n^l\int_0^\infty e^{in\tau A}\,e^{-nA}\,(e^A-1)^{l-1}\,dA=$$

$$=\frac{l}{n}C_n^l\int_0^\infty e^{i\tau a}\,e^{-a}\left(e^{\frac{a}{n}}-1\right)^{l-1}da,$$

după ce am făcut schimbarea de variabilă $nA=a$.
Deci densitatea sa de distribuţie este

$$\rho_{l,n}(a)=C_{n-1}^{l-1}\,e^{-a}\left(e^{\frac{a}{n}}-1\right)^{l-1}$$

sau, folosind formula lui Stirling pentru reprezentarea expresiei C_{n-1}^{l-1} cînd n este mare

$$\rho_{l,n}(a)=\frac{1}{(l-1)}\,e^{-a}\frac{(n-1)^{n-l+\frac{1}{2}}}{(n-l)^{n-l+\frac{1}{2}}}\left[(n-1)\left(e^{\frac{a}{n}}-1\right)\right]^{l-1}.$$

Deci

$$\lim_{n\to\infty}\rho_{l,n}(a)=\frac{1}{(l-1)\,!}\,e^{-a}\,a^{l-1}, \qquad \text{(XXII.1,8)}$$

avînd în vedere că $\lim_{n\to\infty}(n-1)\left(e^{\frac{a}{n}}-1\right)=a.$

Formula este astfel demonstrată.

Teorema 3. *Densitățile de distribuție ale variabilelor* nY_l *și* $n(1-Y_{n-l+1})$ *sînt de asemenea asimptotic euleriene, pentru fiecare* l.

Dacă ținem seamă de (XXII.1,1)

$$P\left(n(1-Y_{n-l+1})<a\right)=P\left(z_l<\ln\frac{1}{1-\frac{a}{n}}\right).$$

Dar, pentru n foarte mare,

$$P\left(z_l<\ln\frac{1}{1-\frac{a}{n}}\right)=P\left(z_l<\frac{a}{n}\right),$$

deoarece

$$\frac{n}{a}\ln\frac{1}{1-\frac{a}{n}}\to 1$$

cînd $n\to\infty$.

Folosind acum formula (XXII.1,8) găsim

$$P\left(n(1-Y_{n-l+1})<a\right)=\frac{1}{(l-1)\,!}\int_0^a e^{-u}\,u^{l-1}\,du. \qquad \text{(XXII.1,9)}$$

Relaţia corespunzătoare variabilei nY_l se stabileşte în acelaşi mod, ţinînd seamă de simetrie şi folosind (XXII.1,8) şi (XXII.1,9)

$$P(nY_l<a)=\frac{1}{(l-1)!}\int_0^a e^{-u}\,u^{l-1}\,du. \qquad \text{(XXII.1,10)}$$

Teorema 4. *Variabilele Y_l şi Y_{n-j+c} sînt asimptotic independente.* Mai precis, ne propunem să demonstrăm următoarea egalitate

$$\lim_{n\to\infty} P\left(Y_l<\frac{a}{n},\quad 1-Y_{n-j+1}<\frac{b}{n}\right)=$$

$$=\frac{1}{(l-1)!\,(j-1)!}\int_0^a\int_0^b u^{l-1}\,v^{j-1}\,e^{-(u+v)}\,du\,dv.$$

In acest scop, observăm cum din (XXII.1,6) rezultă

$$z_{n-j+1}=z_l+\frac{\theta_{l+1}}{n-l}+\ldots+\frac{\theta_{n-j+1}}{j}.$$

Deoarece

$$z_l=\frac{\theta_1}{n}+\frac{\theta_2}{n-1}+\ldots+\frac{\theta_l}{n-l+1},$$

este o variabilă infinitezimală cu $n\to\infty$ în sensul lui Cramér

$$\underset{n\to\infty}{M}[z_l]=0,\quad \underset{n\to\infty}{M}[(z_l-M(z_l))^2]\to 0$$

iar

$$\frac{\theta_{l+1}}{n-l}+\ldots+\frac{\theta_{n-j+1}}{j}-\ln n,$$

este independentă de z_l, rezultă, prin aplicarea propoziţiei 2 a lui Cramér (sfîrşitul acestui capitol), că $z_{n-j+1}-\ln n$ este independentă asimptotic de z_j.
Insă, ţinînd seamă de (XXII.1,1) avem

$$P\left(Y_l<\frac{a}{n},\quad 1-Y_{n-j+1}<\frac{b}{n}\right)=P\left(z_{n-l+1}>\ln\frac{n}{a},\ z_j<\ln\frac{1}{1-\frac{b}{n}}\right)=$$

$$=P\left(z_{n-l+1}-\ln n>-\ln a,\quad z_j<\frac{b}{n}\right)=$$

$$=P(z_{n-l+1}-\ln n>-\ln a)\cdot P\left(z_j<\frac{b}{n}\right).$$

Dar

$$P(z_{n-l+1} - \ln n \geqq -\ln a) = P\left(z_{n-l+1} \geqq \ln \frac{n}{a}\right).$$

Folosind relația (XXII.1,10), avem, însă,

$$P\left(z_{n-l+1} \geqq \ln \frac{n}{a}\right) = P\left(Y_l \leqq \frac{a}{n}\right) = \frac{1}{(l-1)!} \int_0^a u^{l-1} e^{-u}\, du,$$

iar, potrivit cu (XXII.1,8),

$$P\left(z_j < \frac{b}{n}\right) = \frac{1}{(j-1)!} \int_0^b v^{s-1} e^{-v}\, dv;$$

relația (XXII.1,11) se poate considera demonstrată.

§ 2. Aplicație la statistica de ordine

1. *Statistica extensiunii (range).* Considerațiile precedente, valabile pentru variabilele cu distribuție exponențială sau omogenă oarecare, au fost prezentate, fără ca prin aceasta să se micșoreze generalitatea considerațiilor, în special pentru variabilele y și z legate de variabila x prin relațiile (XXII.1,1).

Ele pot deci să fie folosite pentru stabilirea legilor de distribuție ale variabilelor ordonate $X_1, X_2, \ldots, X_n$ în cazul în care $F(x)$ este continuu crescătoare.

Să considerăm, în particular, *extensiunea* $\mathcal{E}_y$ a sistemului $y_1, y_2, \ldots, y_n$. Ea este reprezentată de variabila aleatorie

$$\mathcal{E}_y = Y_n - Y_1.$$

Deoarece, potrivit rezultatelor precedente $n(1-Y_n)$ și nY_1 sînt variabile aleatorii independente, ambele cu densitatea de distribuție limită e^{-a}, $P(n(1-Y_n+Y_1)<a)$ va rezulta din compunerea densității e^{-a} cu ea însăși (teorema 4)

$$\int_0^a e^{-(a-v)} e^{-v}\, dv = ae^{-a}.$$

Prin urmare

$$\lim_{n\to\infty} P\left(1-(Y_n-Y_1)<\frac{a}{n}\right)=\int_0^a ue^{-u}\,du.$$

De unde, ţinînd seamă şi de (XXII.1,1),

$$\lim_{n\to\infty} P\left(1-\frac{b}{n}<F(X_n)-F(X_1)<1-\frac{\alpha}{n}\right)=\int_a^b ue^{-u}\,du. \quad \text{(XXII.2,1)}$$

Probabilitatea extensiunii X_n-X_1 se poate deduce din formula (XXII.2,1) dacă se cunoaşte funcţia $F(a)$, cu ajutorul tabelei de valori ale integralei

$$\int_a^b ue^{-u}\,du=(a+1)\,e^{-a}-(b+1)\,e^{-b}.$$

2. *Statistica quantilelor unui eşantion. Valoarea de rang l a eşantionului $x_1, x_2, .., x_n$, notată de noi cu X_l, este o funcţie aleatorie cu timp discontinuu* ($t=1, 2, \ldots, n, \ldots$). Notaţia ei completă trebuie să fie $X_l(n)$. Pentru fiecare l avem o variabilă aleatorie a rangului l care variază — ca variabilă aleatorie — cu n. Acesta a fost punctul de vedere al teoremelor care preced.

Cu ajutorul acestor funcţii aleatorii se pot defini altele deosebit de importante, cum sînt de pildă *quantilele*, stabilind o relaţie între parametrul de ordine l şi parametrul n de mărime a eşantionului.

Quantila de rang r sau r—quantila este, pentru un n dat, suficient de mare, acea variabilă de ordine $X_l(n)$ pentru care $l=\left[\frac{n}{r}\right]$ (adică l este egal cu cel mai mare întreg cuprins în $\frac{n}{r}$).

Se întrebuinţează denumirea de *rang al quantilei* şi pentru numărul $q=\frac{1}{r}$, pentru care se ia, în general, un număr pozitiv cuprins între 0 şi 1 şi care este mai comod în scris. In acest caz, $l=[nq]$.

Notăm quantila de rang r a eşantionului $x_1, x_2, \ldots, x_n$ cu $Q_r(n)$ avînd deci

$$Q_r(n)=X_{\left[\frac{n}{r}\right]}(n), \quad \text{(XXII.2,2)}$$

r fiind fix. Notăm de asemenea cu Q_r quantila de rang r a distribuţiei $F(a)$, avînd deci $F(Q_r)=\frac{1}{r}$ şi demonstrăm următoarea

Teoremă. *Dacă* $\rho(a)$, *densitatea de distribuție a variabilei* $x=x(\xi)$, *este continuă și strict pozitivă în intervalul* (a_1, a_2) *și dacă* $\left|l-\frac{n}{r}\right|=O(\sqrt{n})$, *atunci*

$$\lim_{n\to\infty} M(Q_r(n))=Q_r, \qquad \text{(XXII.2,3)}$$

iar distribuția asimptotică a variabilei $Q_r(n)$ *în jurul acestei valori medii este cea gaussiană*

$$\lim_{n\to\infty} P\left[\frac{Q_r(n)-Q_r}{\mu_{r,n}}<a\right]=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{a} e^{-\frac{u^2}{2}}du, \qquad \text{(XXII.2,4)}$$

unde $\mu_{r,n}=\frac{1}{\rho(Q_r)}\sqrt{\frac{1}{n}\cdot\frac{1}{r}\left(1-\frac{1}{r}\right)}$ *pentru a cuprins în intervalul* a_1, a_2.

Demonstrația, care urmează folosește faptul că $Z_{n-\left[\frac{n}{r}\right]+1}$ este o sumă de variabile independente și deci are, asimptotic, o distribuție normală în jurul valorii sale medii. In adevăr, după formula (XXII.1,6) avem

$$Z_{n-\left[\frac{n}{r}\right]+1}=\sum_{h=1}^{n-\left[\frac{n}{r}\right]+1}\frac{\theta_h}{n-h+1}.$$

Fiecare dintre variabilele θ_h independente între ele are aceeași funcție de distribuție $1-e^{-x}$ pentru $x>0$ și 0 pentru $x<0$, și avem, în același timp, $M(\theta_h)=1$, $M[(\theta_h-1)^2]=1$.

Să observăm de asemenea că

$$M(|\theta_h-1|^3)=\int_0^{\infty}|u-1|^3 e^{-u}du<\int_0^{1}e^{-u}du+\int_0^{\infty}u^3 e^{-u}du<1.$$

Prin urmare

$$M_n=M\left[Z_{n-\left[\frac{n}{r}\right]+1}\right]=\sum_{h=1}^{n-\left[\frac{n}{r}\right]+1}\frac{1}{n-h+1},$$

$$S_n^2 = M\left[\left(Z_{n-\left[\frac{n}{r}\right]+1} - M_n\right)^2\right] = \sum_{h=1}^{n-\left[\frac{n}{r}\right]+1} \frac{1}{(n-h+1)^2},$$

$$R_n^3 = \sum_{h=1}^{n-\left[\frac{n}{r}\right]+1} M\left[\left|\frac{\theta_h - 1}{n-h+1}\right|^3\right] \leq 5 \sum_{h=1}^{n-\left[\frac{n}{r}\right]+1} \frac{1}{(n-h+1)^3}.$$

Deoarece, cum se observă imediat, condiția lui Liapunov

$$\lim_{n\to\infty} \frac{R_n}{S_n} \to 0$$

este îndeplinită, vom avea

$$\lim P\left\{\frac{1}{S_n}\left(Z_{n-\left[\frac{n}{r}\right]+1} - M_n\right) < a\right\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-\frac{u^2}{2}}\, du.$$

Rămîne să evaluăm mediile M_n și S_n ținînd seamă de expresiile de mai sus. Observăm pentru aceasta că

$$\sum_{k=m_1}^{m_2} \frac{1}{k} = \ln \frac{m_2}{m_1} + \alpha,$$

$$\sum_{k=m_1}^{m_2} \frac{1}{k^2} = \frac{1}{m_1} - \frac{1}{m_2} + \beta.$$

In particular

$$M_n = \ln \frac{n}{\left[\frac{n}{r}\right]} + \alpha_{(n)}$$

$$S_n = \frac{1}{\left[\frac{n}{r}\right]} - \frac{1}{n} + \beta_{(n)},$$

unde

$$|\alpha_{(n)}| < \frac{A}{\left[\frac{n}{r}\right]}, \quad |\beta_{(n)}| < \frac{B}{n^2}.$$

Deci

$$\frac{1}{S_n}\left(Z_{n-\left[\frac{n}{r}\right]+1}-M_n\right)=\frac{Z_{n-\left[\frac{n}{r}\right]+1}-\ln\frac{n}{\left[\frac{n}{r}\right]}}{\sqrt{\frac{n-\left[\frac{n}{r}\right]}{n\left[\frac{n}{r}\right]}}}+\gamma_{(n)}.$$

Insă, ținînd seamă de condiția teoremei, avem

$$\ln\frac{n}{\left[\frac{n}{r}\right]}-\ln r=0\left(\frac{1}{\sqrt{n}}\right).$$

De unde

$$\lim_{n\to\infty}P\left(\frac{Z_{n-\left[\frac{n}{r}\right]+1}-\ln r}{\sqrt{\frac{r-1}{n}}}<a\right)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{a}e^{-\frac{u^2}{2}}du. \qquad \text{(XXII.2,5)}$$

Ne întoarcem acum la quantila de rang r, $X_{\left[\frac{n}{r}\right]}$, determinată de relația

$$Z_{n-\left[\frac{n}{r}\right]+1}=\ln\frac{1}{F\left(X_{\left[\frac{n}{r}\right]}\right)}.$$

Inegalitatea din paranteza egalității (XXII.2,5) se transformă în

$$X_{\left[\frac{n}{r}\right]}>F^{-1}\left(\frac{1}{r}e^{-a\sqrt{\frac{n-1}{r}}}\right),$$

a cărei probabilitate asimptotică este dată de membrul al doilea din (XXII.2,5).

Potrivit formulei ușor de verificat, ținînd seama că $F(x)$ are o derivată $\rho(x)$ pozitivă și continuă

$$F^{-1}\left(\frac{1}{r}e^{-a\sqrt{\frac{n-1}{r}}}\right)=Q_r+\frac{\left(e^{-a\sqrt{\frac{n-1}{r}}}-1\right)}{r\rho(Q_r\delta_n)},$$

iar $\delta_n \to 1$ cînd $n \to \infty$, rezultă imediat egalitatea căutată

$$\lim_{n\to\infty} P\left(\frac{X_{\left[\frac{n}{r}\right]} - Q_r}{\frac{1}{\rho(Q_r)} \sqrt{\frac{1}{n}, \frac{1}{r}\left(1-\frac{1}{r}\right)}} < a \right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-\frac{u^2}{2}} du$$

pentru orice a în intervalul a_1, a_2.

Mediana. *Pentru distribuţia continuă şi crescătoare $F(x)$, mediana este valoarea Q_2 care corespunde valorii* $\frac{1}{2}$

$$F(Q_2) = \frac{1}{2}.$$

In selecţii pare, $X_{\left[\frac{n}{r}\right]}$ există, dar în cele impare această valoare nu există. Ea se înlocuieşte în acest caz cu $X_{\left[\frac{n+1}{2}\right]}$, a cărei distribuţie asimptotică este aceeaşi ca a variabilei $X_{\left[\frac{n}{2}\right]}$.

Dacă distribuţia este simetrică, mediana coincide cu valoarea medie a cărei distribuţie asimptotică este supusă, astfel, de asemenea legii lui Gauss.

§ 3. Cîteva teoreme relative la şirul ordonat $Y_1, Y_2, \ldots, Y_n$, în enunţul şi demonstrarea lui Şerban Gheorghiu

1. *Variabilele $y_1, y_2, \ldots, y_n$ au distribuţia uniformă şi valoarea cuprinsă în intervalul* (0,1).

Introducem variabilele

$$u_1 = Y_1,\ u_2 = Y_2 - Y_1,\ \ldots,\ u_n = Y_n - Y_{n-1} \qquad \text{(XXII.3,1)}$$

şi

$$u_{n+1} = 1 - Y_n. \qquad \text{(XXII.3,2)}$$

Deci

$$u_j \geqq 0,\ \sum_{j=1}^{n+1} u_j = 1. \qquad \text{(XXII.3,3)}$$

In acelaşi timp valoarea funcţiei

$$\Phi(t_1, t_2, \ldots, t_n) = n! \int_{D_n} e^{t_1u_1+t_2u_2+\ldots+t_nu_n}\, du_1 du_2 \ldots du_n, \quad \text{(XXII.3,4)}$$

unde D_n este domeniul definit de inegalităţile

$$u_j \geqq 0, \quad \sum_{j=1}^{n} u_j < 1,$$

iar $t_1, t_2, \ldots, t_n$ sînt reale, este dată de $n!\,[t_1, t_2, \ldots, t_n, 0]$, unde

$$[t_1, t_2, \ldots, t_{n+1}] = \int_{\Delta_{n+1}} e^{t_1u_1+t_2u_2+\ldots+t_{n+1}u_{n+1}}\, du_1 du_2 \ldots du_{n+1}, \quad \text{(XXII.3,5)}$$

Δ_{n+1} fiind domeniul definit de (XXII.3,3), este diferenţa divizată de ordinul n a funcţiei e^t.

Pentru variabila aleatorie

$$\sigma = \sum_{j=1}^{n+1} \lambda_j u_j,$$

unde $\lambda_1, \lambda_2, \ldots, \lambda_{n+1}$ sînt parametri arbitrari, avem

$$\Phi_{\lambda_1, \lambda_2, \ldots, \lambda_{n+1}}(t) = n! \int_{D_n} e^{(\lambda_1u_1+\lambda_2u_2+\ldots+\lambda_{n+1}u_{n+1})\,t}\, du_1 du_2 \ldots du_n =$$

$$= n!\,[\lambda_1 t, \lambda_2 t, \ldots \lambda_{n+1} t]. \quad \text{(XXII.3,6)}$$

2. Dacă toţi λ_j sînt nuli, în afară de $\lambda_h = 1$, formula (XXII.3,6) ne dă funcţia caracteristică laplaciană a variabilei u_h, care după cum se vede imediat, este aceeaşi pentru fiece h.

Avem astfel teorema

Toate variabilele $u_1, u_2, \ldots, u_n$ *au aceeaşi lege de distribuţie.*

Folosind o formulă cunoscută la reprezentarea diferenţelor divizate se găseşte imediat că

$$\Phi_h(t) = \frac{n!}{t^n}\left[e^t - 1 - \frac{t}{1!} - \ldots - \frac{t^{n-1}}{(n-1)!}\right].$$

De aici rezultă densitatea de distribuţie a fiecăreia dintre variabilele u_j

$$\rho_{h,n}(x)=\begin{cases} 0 & \text{pentru } x<0,\ x>1 \\ n(1-x)^{n-1} & \text{pentru } 0<x<1, \end{cases}$$

ceea ce dă pentru funcţia de distribuţie

$$F_{h,n}(a)=\begin{cases} 0 & , \quad a<0, \\ 1-(1-a)^n & , \quad 0<a<1, \\ 1 & , \quad a>1. \end{cases}$$

3. *Funcţia de distribuţie a unei sume de k variabile u_j este evident independentă de indicii acestor variabile.* Se iau pentru aceasta primele k constante $\lambda_1, \lambda_2, \ldots, \lambda_k$ egale cu unu şi celelalte egale cu zero. Funcţia caracteristică laplaciană ce se obţine este

$$\Phi_n^{(k)}(t)=n!\,[t,\ t, \ldots, t,\ 0,\ 0, \ldots, 0]$$

şi folosind aceeaşi formulă de reprezentare a diferenţelor divizate avem a doua teoremă

$$\Phi_n^{(k)}(t)=\frac{n!}{2\pi i}\int_C \frac{e^z dz}{z^{n+1-k}(z-t)^k}.$$

Din aceasta se obţine prin transformări convenabile

$$\Phi_n^{(k)}(t)=\frac{n!}{(n-k)!\,(k-1)!}\int_0^1 u^{k-1}(1-u)^{n-k}e^{tu}\,du,$$

ceea ce dă imediat pentru momentele variabilei $u_1+u_2+\ldots+u_k$ valorile

$$M_{n,l}^{(k)}=\frac{n!}{(n-k)!\,(k-1)!}\cdot\frac{\Gamma(k+l)\,\Gamma(n-k+1)}{(n+l+1)}.$$

Funcţia de distribuţie este prin urmare

$$F_n^{(k)}(a)=\begin{cases} 0 & , \quad x<0 \\ \sum\limits_{j\geq k} C_n^j a^j (1-a)^{n-j} & , \quad 0<x<1 \\ 1 & , \quad 1<x. \end{cases}$$

4. *Cazul general.* Funcţia caracteristică laplaciană a sumei $\sigma = \sum_{j=1}^{n+1} \lambda_j u_j$ se obţine prin aplicarea aceleiaşi formule generale de reprezentare a diferenţiei generalizate, ceea ce dă

$$\Phi^n_{\lambda_1, \lambda_2, \ldots, \lambda_{n+1}}(t) =$$

$$= \frac{n!}{2\pi t^n i} \int_C \frac{e^{tz} dz}{(z-\lambda_1')^{\mu_1}(z-\lambda_2')^{\mu_2} \ldots (z-\lambda_h')^{\mu_h}},$$

unde $\lambda_1', \lambda_2', \ldots, \lambda_h'$ sînt cele h valori distincte ale parametrilor λ_j şi $\mu_1, \mu_2, \ldots, \mu_h$ sînt ordinele lor respective de multiplicitate, iar C este un cerc care cuprinde în interior punctele $\lambda_1', \lambda_2', \ldots, \lambda_h'$. Aceasta conduce pînă la urmă la o expresie de forma

$$\frac{n!}{t^n} \sum P_j(t) e^{\lambda_j' t},$$

unde $P_j(t) = a_{0,j} + a_{1,j} t + \ldots + a_{\mu j - 1, j} t^{\mu j - 1}$.

5. Dacă ordonăm şirul $u_1, u_2, \ldots, u_{n+1}$, obţinem un nou şir

$$U_1, U_2, \ldots, U_{n+1}$$

căruia îi asociem variabilele $v_1, v_2, \ldots, v_{n+1}$ definite precum urmează

$$U_1 = \frac{v_1}{n+1},$$

$$U_2 = \frac{v_1}{n+1} + \frac{v_2}{n},$$

. . . . . .

$$U_k = \frac{v_1}{n+1} + \frac{v_2}{n} + \ldots + \frac{v_k}{n+2-k},$$

. . . . . .

$$U_{n+1} = \frac{v_1}{n+1} + \frac{v_2}{n} + \ldots + \frac{v_{n+1}}{1}.$$

Deoarece

$$\frac{D(U_1, U_2, \ldots, U_n)}{D(v_1, v_2, \ldots, v_n)} = \frac{1}{(n+1)!},$$

funcţia caracteristică laplaciană relativă la distribuţia definită de variabilele $v_1, v_2, \ldots, v_n$ este

$$n! \int_{D_n(v)} e^{t_1 v_1 + t_2 v_2 + \ldots + t_n v_n} \, dv_1 \, dv_2 \ldots dv_n,$$

domeniul $D_n(v)$ fiind caracterizat de inegalităţile

$$v_j \geqq 0, \quad \sum_{j=1}^{n} v_j < 1.$$

Aceasta arată însă, ca şi în cazul precedent, că *variabilele v_j ale statisticii de speţa a doua au aceeaşi distribuţie oricare ar fi j şi anume aceeaşi distribuţie ca şi variabilele u_j.*

Teorema este valabilă pentru toate statisticile de ordin superior, ceea ce este de altfel de aşteptat, dată fiind forma generală a rezultatelor obţinute pentru statisticile de ordinul întîi.

Capitolul XXIII

Verificarea ipotezelor

§ 1. Consideraţii generale

Să considerăm o distribuţie unidimensională, a cărei funcţie de frecvenţă

$$f(x; \alpha_1, \alpha_2, \ldots, \alpha_k)$$

depinde de k parametri

$$\alpha_1, \alpha_2, \ldots, \alpha_k,$$

care presupunem că reprezintă coordonatele unui punct

$$\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_k)$$

într-un spaţiu euclidian A cu k dimensiuni.

Fie B o mulţime boreliană din spaţiul euclidian k-dimensional A. Dacă punctul $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_k)$ aparţine lui B, sîntem în prezenţa unei ipoteze pe care o notăm prin

$$H_0 : \alpha \in B.$$

Spunem că $H_0 : \alpha \in B$ reprezintă o ipoteză, fiindcă punctul α, deşi are o valoare bine determinată, totuşi această valoare este pentru noi necunoscută. Apartenenţa punctului α mulţimii B poate să aibă loc sau nu în realitate. De aceea, aceasta constituie o ipoteză H_0, care urmează să fie verificată. O asemenea ipoteză este de multe ori denumită *ipoteza nulă*.

Ipotezele de care ne ocupăm sînt, prin urmare, destul de restrictive, deoarece ele nu se referă la totalitatea afirmaţiilor care se pot face în legătură cu o distribuţie statistică, ci numai la afirmaţiile referitoare la valorile numerice pe care le pot lua parametrii dintr-o distribuţie statistică specificată. Totuşi, în aplicaţiile practice, astfel de ipoteze sînt deseori întîlnite şi modul lor de verificare prezintă un deosebit interes.

Dacă mulţimea boreliană B se reduce la un punct

$$\alpha^0 = (\alpha_1^0, \alpha_2^0, \ldots, \alpha_k^0)$$

avem de verificat ipoteza

$$H_0 : \alpha = \alpha^0$$

sau

$$\alpha_1 = \alpha_1^0, \alpha_2 = \alpha_2^0, \ldots, \alpha_k = \alpha_k^0.$$

In acest caz, ipoteza este *simplă* şi, după cum se vede, fiecare parametru nu poate lua decît o singură valoare bine determinată. In cazul cînd mulţimea boreliană B nu se reduce la un punct, avem o ipoteză *compusă*.

Teoria testelor de verificare a ipotezelor a fost creată de I. Neyman şi E. Pearson.

§ 2. Modul de verificare a ipotezelor

Verificarea ipotezelor se face pe baza unei selecţii de n variabile

$$x_1, x_2, \ldots, x_n$$

pe care le considerăm drept coordonatele unui punct

$$x = (x_1, x_2, \ldots, x_n)$$

dintr-un spaţiu euclidian cu n dimensiuni R_n.

Fie

$$H : \alpha \in B$$

ipoteza pe care vrem s-o verificăm. Dacă există o regiune U, numită *regiune critică*, în spaţiul de selecţie R_n, astfel încît

$$P(x \in U \mid \alpha \subset B) = \delta,$$

unde δ este un număr ales de noi foarte apropiat de zero, respingem ipoteza H_0 cînd, în urma unei selecţii efectuate, punctul $x = (x_1, x_2, \ldots, x_n)$ cade în U şi admitem ipoteza H_0 cînd punctul $x = (x_1, x_2, \ldots, x_n)$ cade în afara lui U.

Este evident că regiunea critică U nu este determinată în mod unic. Oricare dintre aceste regiuni U ne poate da un procedeu pentru verificarea ipotezei H_0: $\alpha \in B$, care poartă numele de *test*. Există, deci, pentru verificarea aceleiaşi ipoteze mai multe teste, dintre care este preferabil să alegem pe cel mai avantajos, dacă există. Pentru a înţelege modul de aplicare a unui test vom raţiona pe un exemplu.

Să considerăm, astfel, distribuţia normală $N(m, \sigma)$. Ne propunem să verificăm ipoteza

$$H_0 : m = m_0,$$

m_0 fiind un număr dat.

Spaţiul $A(m, \sigma)$ al parametrilor este în cazul de faţă un plan. In acest caz, ipoteza $H_0 : m = m_0$ înseamnă că punctul reprezentativ al parametrilor

$$\alpha = (m, \sigma)$$

se găseşte pe o dreaptă. Avem, deci, de verificat o ipoteză compusă.

Dacă distribuţia este normală, $N(m, \sigma)$, am arătat în paragraful precedent că putem să scriem

$$P\left(-t \leq \sqrt{n}\,\frac{\bar{x} - m}{\sqrt{\overline{m}}} \leq t\right) = \frac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{\pi(n-1)}\,\Gamma\left(\frac{n-1}{2}\right)} \int_{-t}^{t} \left(1 + \frac{u^2}{n-1}\right)^{-\frac{n}{2}} du,$$

cu alte cuvinte, dacă ipoteza $H_0 : m = m_0$ este adevărată, trebuie să aibă loc egalitatea

$$P\left(-t \leq \sqrt{n}\,\frac{\bar{x} - m_0}{\sqrt{m_2}} \leq t\right) = \frac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{\pi(n-1)}\,\Gamma\left(\frac{n-1}{2}\right)} \int_{-t}^{t} \left(1 + \frac{u^2}{n-1}\right)^{-\frac{n}{2}} du.$$

Scanned with OKEN Scanner

Alegîndu-ne, atunci, un număr δ foarte apropiat de zero, putem determina un alt număr t_δ, astfel ca

$$P\left(-t_\delta \leqslant \sqrt{n}\,\frac{\bar{x}-m_0}{\sqrt{m_2}} \leqslant t_\delta\right) = 1-\delta.$$

In acest mod, am definit în spaţiul selecţiei R_n o regiune

$$U\left(\sqrt{n}\,\frac{\bar{x}-m_0}{\sqrt{m_2}} \geqslant t_\delta, \quad \sqrt{n}\,\frac{\bar{x}-m_0}{\sqrt{m_2}} \leqslant -t_\delta\right).$$

Dacă, efectuînd o selecţie de mărime n, punctul rezultat $x=(x_1, x_2, \ldots x_n)$ cade în regiunea U, respingem ipoteza $H_0 : m = m_0$ ca nejustă cu pragul de semnificaţie δ. Dacă însă punctul $x=(x_1, x_2, \ldots, x_n)$ cade în afara regiunii U, admitem ipoteza $H_0 : m=m_0$ cu pragul de semnificaţie δ.

Este de observat apropierea metodelor de verificare a testelor şi a domeniilor de încredere. Totuşi între ele există o deosebire esenţială. La metoda verificării ipotezelor se foloseşte probabilitatea referitoare la un domeniu din spaţiul selecţiilor, pe cînd la metoda domeniilor de încredere se folosesc probabilităţile regiunilor referitoare la un domeniu din spaţiul parametrilor.

§ 3. Erori de genul întîi şi al doilea

Din cele arătate mai înainte rezultă că admiterea sau respingerea unei ipoteze este o convenţie pe care o facem. Intr-un număr imens de cazuri, punctul $x=(x_1, x_2, \ldots, x_n)$, rezultat al unei selecţii, cade în întreg spaţiul R_n al acestei selecţii. In baza legii numerelor mari, numai într-un număr foarte mic de cazuri va cădea în regiunea critică U. In medie, dacă am făcut N selecţii, numai $N\delta$ dintre ele au căzut în U, iar în imensa lor majoritate au căzut în afară de U ($\delta \approx 0$).

Considerînd că într-o singură experienţă căderea lui x în U este un fenomen rar care se întîmplă numai în $\delta^0/_0$ cazuri, respingem ipoteza dacă acest lucru se întîmplă. In cazul acesta, ştim de la început că nu este exclus să facem o eroare şi anume, ca punctul $x=(x_1, x_2, \ldots, x_n)$ să cadă în regiunea critică U, ipoteza H_0 fiind justă. Prin urmare, putem elimina o ipoteză justă cu probabilitatea δ de a se întîmpla aceasta. Eroarea pe care o facem, în acest caz, este o *eroare de genul întîi.*

Dar mai putem face şi alt gen de erori. Se poate întîmpla ca punctul $x=(x_1, x_2, \ldots, x_n)$ să cadă în afara regiunii critice U. Deşi atunci admitem ipoteza H_0, totuşi nu este exclus ca ipoteza H_0 să fie nejustă. Deci putem face eroarea de a admite o ipoteză falsă. Eroarea pe care o facem de data aceasta este o *eroare de genul al doilea.*

Presupunînd ipoteza justă

$$H_0=\alpha\in B$$

şi ipoteza falsă

$$H_0=\alpha\in\overline{B},$$

$\overline{B}$ fiind complementara lui B,
avem de considerat următoarele probabilităţi:

$$P(x\in U \mid \alpha\in B), \qquad \text{(XXIII.3,1)}$$

$$P(x\in U \mid \alpha\in B), \qquad \text{(XXIII.3,2)}$$

$$P(x\in U \mid \alpha\in B), \qquad \text{(XXIII.3,3)}$$

şi

$$P(x\in\overline{U} \mid \alpha\in B), \qquad \text{(XXIII.3,4)}$$

unde $\overline{U}$ este complementara lui U.

Probabilitatea (XXIII.3,1) reprezintă probabilitatea ca punctul x să cadă în regiunea critică U, în ipoteza că H_0 este justă. Prin urmare, reprezintă probabilitatea de a respinge o ipoteză justă.

Probabilitatea (XXIII.3,2) reprezintă probabilitatea ca punctul x să cadă de asemenea în regiunea critică U, ipoteza H_0 fiind falsă. Deci reprezintă probabilitatea de a respinge o ipoteză falsă.

Probabilitatea (XXIII.3,3) reprezintă probabilitatea ca punctul x să cadă în afara regiunii critice U, în cazul cînd ipoteza H_0 este justă. Cu alte cuvinte, este probabilitatea de a admite o ipoteză justă.

In sfîrşit, probabilitatea (XXIII.3,4) reprezintă probabilitatea ca punctul x să cadă în afara regiunii critice U, ipoteza H_0 fiind falsă. Avem astfel probabilitatea de admite o ipoteză falsă.

Cu cît probabilităţile (XXIII.3,1) şi (XXIII.3,4) sînt mai mici, cu atît testul este evident mai puternic.

După cum se observă, probabilitatea (XXIII.3,1) este probabilitatea de a comite o eroare de genul întîi, adică de a respinge o ipoteză justă, iar probabilitatea (XXIII.3,4) înseamnă probabilitatea de a comite o eroare de genul al doilea, adică de a admite o ipoteză falsă.

Scanned with OKEN Scanner

Alegînd un test, noi putem, prin mărirea volumului selecţiei, să facem ca probabilitatea comiterii unei erori de un anumit gen să fie arbitrar de mică. Nu putem însă să facem acelaşi lucru pentru probabilităţile erorilor de ambele genuri. Alegîndu-ne un test, să facem volumul selecţiei atît de mare, încît probabilitatea ca să obţinem o eroare de genul întîi, să fie, de pildă, egală cu 0,05. Atunci probabilitatea comiterii unei erori de genul al doilea rezultă ca o consecinţă, ea nu mai stă la libera noastră dispoziţie. Invers, putem fixa probabilitatea comiterii unei erori de genul al doilea şi atunci rezultă ca o consecinţă probabilitatea comiterii unei erori de genul întîi. Nu se poate afirma care dintre aceste probabilităţi trebuie să fie mai mică sau mai mare, deoarece nu există o regulă generală în această privinţă.

De exemplu, dacă dorim să verificăm componenţa unui medicament, care peste un anumit grad de concentrare devine vătămător, comiterea unei erori de genul al doilea este mult mai primejdioasă decît comiterea unei erori de genul întîi. In schimb, la acceptarea unui articol de îmbrăcăminte, care nu corespunde întocmai condiţiilor de fabricaţie, eroarea de primul gen este mai gravă decît eroarea de genul al doilea.

§ 4. Puterea unui test

Prin definiţie, *puterea unui test* este probabilitatea de a respinge ipoteza. Aceasta înseamnă ca punctul $x=(x_1, x_2, \ldots, x_n)$ rezultat din selecţie să cadă în regiunea critică U.

De unde rezultă că, dacă ipoteza H_0 este justă, puterea testului trebuie să fie cît mai mică, iar dacă ipoteza H_0 este falsă, puterea testului trebuie să fie cît mai mare.

Să arătăm modul cum variază puterea testelor unei ipoteze H_0, considerînd următorul exemplu.

Fie distribuţia normală $N(m, \sigma)$ a cărei funcţie de frecvenţă este

$$\frac{1}{\sqrt{2\pi}\cdot\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}},$$

unde σ este dat şi fie, de asemenea, ipoteza simplă

$$H_0 : m = m_0.$$

Deoarece se ştie că funcţia de selecţie

$$\sqrt{n}\,\frac{\bar{x}-m}{\sigma}$$

urmează legea normală $N(0,1)$ putem determina două numere α şi β astfel ca

$$P\left(\alpha \leqslant \sqrt{n}\,\frac{\bar{x}-m}{\sigma} \leqslant \beta\right) = \frac{1}{\sqrt{2\pi}}\int_{\alpha}^{\beta} e^{-\frac{t^2}{2}}dt = 1-\delta, \qquad \text{(XXIII.4,1)}$$

δ fiind un număr foarte apropiat de zero.
Dacă ipoteza $H_0: m=m_0$ este adevărată, trebuie să avem

$$P\left(\alpha \leqslant \sqrt{n}\,\frac{\bar{x}-m}{\sigma} \leqslant \beta\right) = \frac{1}{\sqrt{2\pi}}\int_{\alpha}^{\beta} e^{-\frac{t^2}{2}}\,dt$$

şi deci regiunea critică în spaţiul selecţiei R_n este

$$U\left(\bar{x} \leqslant m_0 + \frac{\alpha\sigma}{\sqrt{n}},\ \bar{x} \geqslant m_0 + \frac{\beta\sigma}{\sqrt{n}}\right), \qquad \text{(XXIII.4,2)}$$

Observăm că regiunea critică este formată prin împărţirea spaţiului R_n în subspaţii, prin hiperplanele

$$x_1+x_2+\ldots+x_n = m_0+\frac{\alpha\sigma}{\sqrt{n}},$$

şi

$$x_1+x_2+\ldots+x_n = m_0+\frac{\beta\sigma}{\sqrt{n}}.$$

Deoarece numerele α şi β sînt determinate prin unica relaţie

$$\frac{1}{\sqrt{2\pi}}\int_{\alpha}^{\beta} e^{-\frac{t^2}{2}}\,dt = 1-\delta,$$

există o infinitate de teste corespunzătoare intervalelor $[\alpha, \beta]$. Toate aceste teste au pragul de semnificaţie δ, adică

$$P(x \in \overline{U}) = \delta.$$

Alegîndu-ne unul dintre aceste teste (α, β), să calculăm puterea lui, condiţionată de faptul că valoarea mediei teoretice este egală cu m_1.

Scanned with OKEN Scanner

Prin definiţie, puterea testului este probabilitatea ca punctul x să cadă în regiunea critică U, adică

$$P(x \in U \mid m = m_1) = \pi(m_1).$$

Din egalitatea

$$P(x \in \overline{U} \mid m = m_1) + P(x \in U \mid m = m_1) = 1,$$

deducem

$$\pi(m_1) = 1 - \frac{\sqrt{n}}{\sqrt{2\pi}} \int_{m_0 + \frac{\alpha\sigma}{\sqrt{n}}}^{m_0 + \frac{\beta\sigma}{\sqrt{n}}} e^{-\frac{n(t-m_1)^2}{2\sigma^2}} dt,$$

avînd în vedere (XXIII.4,2) şi cunoscînd că funcţia de selecţie $\overline{x}$ se distribuie normal, $N\left(m, \frac{\sigma}{\sqrt{n}}\right)$.

Efectuînd substituţia

$$\frac{\sqrt{n}\,(t-m_1)}{\sigma} = \xi,$$

obţinem

$$\pi(m_1) = 1 - \frac{1}{\sqrt{2\pi}} \int_{\alpha + \frac{\sqrt{n}}{\sigma}(m_0 - m_1)}^{\beta + \frac{\sqrt{n}}{\sigma}(m_0 - m_1)} e^{-\frac{\xi^2}{2}} d\xi. \qquad \text{(XXIII.4,3)}$$

Se vede că, oricare ar fi testul (α, β), funcţia $\pi(m_1)$, care se mai numeşte *funcţie de putere*, ia în punctul m_0 valoarea

$$\pi(m_0) = \delta,$$

ţinînd seamă de relaţia (XXIII.4,1).

Să considerăm două teste

$$(\alpha, \beta), \quad (\overline{\alpha}, \overline{\beta}),$$

ale căror funcţii de putere sînt respectiv

$$\pi(m_1), \quad \overline{\pi}(m_1).$$

Dacă are loc inegalitatea

$$\pi(m_1) \geqq \overline{\pi}(m_1),$$

este evident că testul (α, β) este preferabil testului $(\overline{\alpha}, \overline{\beta})$, deoarece în cazul nostru puterea reprezintă probabilitatea de a respinge o

ipoteză falsă. Cu cît această probabilitate este mai mare, cu atît testul este mai bun.

In general, nu vom avea îndeplinită inegalitatea de mai sus, pentru toate valorile posibile ale lui m_1. Cu alte cuvinte, numai în cazuri cu totul particulare putem găsi un test care să fie mai bun decît celelalte. Vom putea determina însă teste care, pentru anumite intervale de variaţie ale lui m_1, să fie mai puternice decît toate celelalte.

In acest scop, să studiem comportarea funcţiei de putere $\pi(m_1)$ pentru diferitele valori ale lui α şi β. Derivata funcţiei de putere (XXIII.4,3) este

$$\pi'(m_1)=\frac{\sqrt{n}}{\sqrt{2\pi}\cdot\sigma}\left[e^{-\frac{1}{2}\left[\beta+\frac{\sqrt{n}}{\sigma}(m_0-m_1)\right]^2}-e^{-\frac{1}{2}\left[\alpha+\frac{\sqrt{n}}{\sigma}(m_0-m_1)\right]^2}\right].$$

Această derivată se anulează pentru

$$m_1=m_0+\frac{\sigma}{2\sqrt{n}}(\alpha+\beta).$$

Calcule elementare ne arată că

$$\pi'(m_1)<0 \qquad \left(m_1<m_0+\frac{\sigma}{2\sqrt{n}}(\alpha+\beta)\right),$$

$$\pi'(m_1)=0 \qquad \left(m_1=m_0+\frac{\sigma}{2\sqrt{n}}(\alpha+\beta)\right),$$

$$\pi'(m_1)>0 \qquad \left(m_1>m_0+\frac{\sigma}{2\sqrt{n}}(\alpha+\beta)\right).$$

Prin urmare funcţia de putere $\pi(m_1)$ a testului (α, β) admite un minim în punctul $m_1=m_0+\frac{\sigma}{2\sqrt{n}}(\alpha+\beta)$ fiind descrescătoare de la $-\infty$ pînă la m_1 şi crescătoare de la m_1 pînă la ∞.

In particular dacă $\alpha+\beta=0$, funcţia de putere $\pi(m_1)$ admite un minim în punctul $m_1=m_0$.

Dacă $\alpha=-\infty$, minimul funcţiei de putere $\pi(m_1)$ se obţine pentru $m_1=-\infty$ şi atunci funcţia $\pi(m_1)$ este continuu crescătoare de la zero la unu, cînd m_1 variază de la $-\infty$ la $+\infty$.

Dacă $\beta=+\infty$, minimul funcţiei de putere $\pi(m_1)$ se obţine pentru $m_1=+\infty$ şi funcţia $\pi(m_1)$ este continuu descrescătoare de la unu la zero, cînd m_1 variază de la $-\infty$ la $+\infty$.

Scanned with OKEN Scanner

Intre aceste două curbe extreme se situează toate celelalte reprezentări grafice ale funcţiei de putere. Astfel,
dacă $\alpha+\beta<0$, minimul funcţiei de putere este la stinga lui m_0,
dacă $\alpha+\beta>0$, minimul funcţiei de putere este la dreapta lui m_0.
Arătăm mai jos reprezentările grafice ale funcţiei de putere $\pi(m_1)$ pentru cazurile examinate (fig. XXIII.4,1).

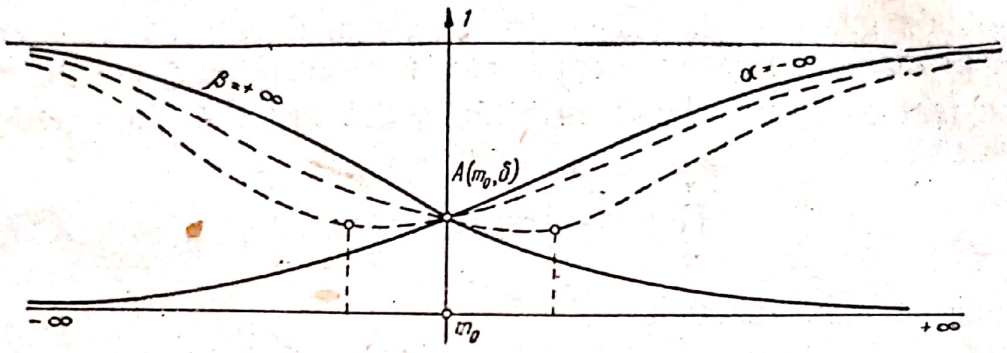


Fig. XXIII.4,1

După cum reiese din grafic, nu există un test care pentru toate valorile lui m_1 de la $-\infty$ la $+\infty$ să ia valori mai mari decît celelalte. Pe porţiuni de variaţie ale lui m_1, putem găsi însă teste mai puternice.

Astfel, pentru toate valorile $m_1>m_0$, testul corespunzător lui $\alpha=-\infty$ este cel mai bun. Cu alte cuvinte, dintre toate testele care ne dau probabilitatea de a comite o eroare de genul întîi, în cazul cînd considerăm că valorile adevărate ale parametrului ar fi mai mari decît m_0, testul corespunzător lui $\alpha=-\infty$ ne dă cea mai mică probabilitate de a comite o eroare de genul al doilea.

In mod analog, testul corespunzător lui $\beta=+\infty$ este cel mai bun, cînd considerăm valorile $m_1<m_0$.

§ 5. Examinarea cazului unui parametru unic care ia două valori

Cazul cel mai simplu este evident cînd funcţia de frecvenţă a distribuţiei teoretice are un singur parametru, care poate lua numai două valori. In acest caz numai două ipoteze sînt posibile. Modul de dependenţă al probabilităţilor de comitere a celor două feluri de erori poate fi în acest caz analizat pe larg.

Fie $f(x;\alpha)$ funcţia de frecvenţă a distribuţiei teoretice. α poate lua numai valorile α_0 şi α_1. Ne propunem să verificăm ipoteza H_0

$$H_0 : \alpha=\alpha_0.$$

In acest scop, vom arăta că putem găsi un test care să fie cel mai avantajos dintre toate, în sensul că, dacă luăm în considerare toate testele admiţînd aceeaşi probabilitate δ pentru a comite o eroare de genul întîi, există printre ele unul care admite cea mai mică probabilitate pentru comiterea unei erori de genul al doilea. Acest test este determinat prin domeniul critic A din spaţiul de selecţie

$$\prod_{i=1}^{n} f(x_i;\alpha_1)\geqq c\prod_{i=1}^{n} f(x_i;\alpha_0), \qquad \text{(XXIII.5,1)}$$

c rezultînd din egalitatea

$$P[x\in A\,|\,\alpha_0]=\delta, \qquad \text{(XXIII.5,2)}$$

cu condiţia ca $f(x;\alpha)$ să fie mărginită.

Vom arăta acest lucru în modul următor. Fie B domeniul critic corespunzător unui alt test. Avem

$$P(x\in B\,|\,\alpha_0)=\delta. \qquad \text{(XXIII.5,3)}$$

Pe de altă parte

$$\begin{aligned} P[x\in(A-A\cap B)\,|\,\alpha_0]&=P(x\in A\,|\,\alpha_0)-P(x\in A\cap B\,|\,\alpha_0)=\\ =\delta-P[x\in A\cap B\,|\,\alpha_0]&=P(x\in B\,|\,\alpha_0)-P(x\in A\cap B\,|\,\alpha_0)=\\ &=P[x\in(B-A\cap B)\,|\,\alpha_1), \qquad \text{(XXIII.5,4)}\end{aligned}$$

$$P(x\in(A-A\cap B)\,|\,\alpha_1)\geqq cP[x\in(A-A\cap B\,|\,\alpha_0]. \qquad \text{(XXIII.5,5)}$$

Ultima inegalitate este justificată prin faptul că

$$P[x\in(A-A\cap B)\,|\,\alpha_1]=\iint\ldots\int_{(A-A\cap B)}\prod_{i=1}^{n} f(x_i;\alpha_1)\,dx_i,$$

$$P[x\in(A-A\cap B)\,|\,\alpha_0]=\iint\ldots\int_{(A-A\cap B)}\prod_{i=1}^{n} f(x_i;\alpha_0)\,dx_i,$$

$$\iint\ldots\int_{(A-A\cap B)}\prod_{i=1}^{n} f(x_i;\alpha_1)\,dx_i\geqq c\iint\ldots\int_{(A-A\cap B)}\prod_{i=1}^{n} f(x_i;\alpha_0)\,dx_i,$$

în baza inegalităţii (XXIII.5,1).

Mai departe, din relaţia (XXIII.5,4) deducem

$$cP[x \in (A-A\cap B) \mid \alpha_0] = cP[x \in (B-A\cap B) \mid \alpha_0]$$

şi

$$cP[x \in (B-A\cap B) \mid \alpha_0] \geqq P[x \in (B-A\cap B) \mid \alpha_0],$$

deoarece

$$c \int\int \dots \int_{(B-A\cap B)} \prod_{i=1}^{n} f(x_i;\alpha)\, dx_i \geqq$$

$$\geqq \int\int \dots \int_{(B-A\cap B)} \prod_{i=1}^{n} f(x_i;\alpha)\, dx_i, \qquad \text{(XXIII.5,6)}$$

întrucît în domeniul complementar lui A avem, în conformitate cu (XXIII.5,1)

$$c \prod_{i=1}^{n} f(x_i;\alpha_0) \geqq \prod_{i=1}^{n} f(x_i;\alpha_0).$$

Ţinînd seamă de (XXIII.5,5) şi (XXIII.5,6), obţinem

$$P[x \in (A-A\cap B) \mid \alpha_1] \geqq P[x \in (B-A\cap B) \mid \alpha_1]$$

sau

$$P[x \in A \mid \alpha_1] \geqq P(x \in B \mid \alpha_1).$$

Prin urmare, puterea testului avînd domeniul critic A este mai mare decît puterea testului cu domeniul critic B. Rămîne numai să demonstrăm existenţa domeniului critic A.

Să notăm

$$\psi(c) = P[x \in A \mid \alpha_0].$$

Este evident că, dacă $c=0$, avem $\psi(c)=1$, deoarece domeniul A devine atunci

$$\prod_{i=1}^{n} f(x_i;\alpha_1) \geqq 0,$$

adică întreg spaţiul R_n.

Pe de altă parte, dacă c creşte, domeniul A descreşte şi deci $\psi(c)$ este funcţie necrescătoare de c.

In baza inegalităţii (XXIII.5,1) avem în A

$$\int\int \dots \int_{A} \prod_{i=1}^{n} f(x_i;\alpha_1)\, dx_i \geqq c \int\int \dots \int_{A} \prod_{i=1}^{n} f(x_i;\alpha_0)\, dx_i.$$

Membrul întîi al acestei inegalităţi reprezentînd o probabilitate este cel mult egal cu unu. Avem deci

$$c\psi(c)\leqq 1.$$

Rezultă astfel

$$0\leqq\psi(c)\leqq\frac{1}{c}$$

şi deci, cînd $c\to\infty$, avem $\psi(c)\to 0$.

Egalitatea

$$\psi(c)=\delta \qquad (0\leqq\delta\leqq 1)$$

are atunci totdeauna loc pentru o valoare unică determinată a lui c. S-ar putea prezenta un caz de excepţie numai dacă în punctul c, soluţie a ecuaţiei precedente, există un punct de discontinuitate, deoarece în acest caz în punctul c valoarea funcţiei $\psi(c)$ nu ar fi determinată. Observăm însă că, $\psi(c+0)-\psi(c-0)$ reprezintă probabilitatea ca punctul x să cadă în regiunea

$$\prod_{i=1}^{n} f(x_i;\alpha_1)=c\prod_{i=1}^{n} f(x_i;\alpha_0).$$

Eliminînd această mulţime din A şi adăugînd-o la complementara lui A, discontinuitatea din punctul c dispare şi deci problema este rezolvată.

O aplicaţie a acestei proprietăţi se poate da în cazul clasic al distribuţiei normale.

Să presupunem că distribuţia teoretică este $N(m, 1)$ şi că referitor la m putem face două ipoteze

$$H_0 : m=m_0, \quad H_1 : m=m_1.$$

Ne propunem să determinăm cel mai bun test pentru verificarea lui H_0.

Regiunea critică pentru acest test este dată de (XXIII.5,1), care în acest caz devine

$$e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-m_1)^2}\geqq ce^{-\frac{1}{2}\sum(x_i-m_0)^2},$$

care se poate pune sub forma

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(x_i-m_0)\geqq\frac{\ln c}{(m_1-m_0)\sqrt{n}}+\frac{\sqrt{n}}{2}(m_1-m_0).$$

Trebuie să determinăm pe c, astfel ca probabilitatea acestei inegalități să fie δ.

Variabila

$$\frac{1}{\sqrt{n}}\left(\sum_{i=1}^{n} x_i - m_0\right)$$

este distribuită normal $N(0, 1)$.

Notînd

$$h = \frac{\ln c}{(m_1 - m_0)\sqrt{n}} = \frac{\sqrt{n}}{2}(m_1 - m_0),$$

putem determina un h_δ, astfel ca

$$P\left[\frac{1}{\sqrt{n}}\left(\sum_{i=1}^{n} x_i - m_0\right) \geqq h_\delta\right] = \delta,$$

δ fiind un număr ales, apropiat de 1.

Am determinat astfel domeniul critic care, după cum se vede, nu depinde de m_2.

BIBLIOGRAFIE

1. *Herald Cramér.* Mathematical methods of statistics. Princeton, 1946.
2. *M. G. Kendall.* The advanced theory of statistics. London, 1948.
3. *L. Schmetterer.* Einführung in die Mathematische Statistik. Wien, 1956.
4. *S. S. Wilks.* Mathematical statistics. Princeton, 1950.
5. *W. E. Deming.* Some theory of sampling. New-York, 1950.
6. *Б. В. Гнеденко.* Курс теории вероятностей. Moscova, 1954.
7. *H. Gebelein.* Zahlung wirklichkeit. Heidelberg, 1949.
8. *C. R. Rao.* Advanced statistical methods in biometric research. New-York, 1952.
9. *I. V. Dunin-Barkovski* și *N. V. Smirnov.* Teoria probabilităților și statistica matematică în tehnică (curs litografiat). București, 1956.
10. *R. A. Fischer.* Les méthodes statistiques adaptées à la recherche scientifique. Paris, 1937.
11. *G. Darmois.* Leçons sur l'éstimation statistique. Paris, 1947.
12. *M. Hausen, W. Hurwitz, W. Madow.* Sample survey methods and theory. New-York, 1953.
13. *R. Fortet.* Calcul des probabilités. Paris, 1950.
14. *O. Onicescu, G. Mihoc, C. T. Ionescu-Tulcea.* Calculul probabilităților și aplicații. București, 1956.
15. *O. Onicescu.* Calculul probabilităților, Editura Tehnică, București, 1956.
16. *Alfréd Rényi.* On the theory of order Statistics, Acta Mathematica, Academiae Scientiarum Hungaricae, Tom IV, Fasciculi 3—4, 1953 cu o bogată bibliografie.
17. *Н. В. Смирнов.* О статической оценке переходных вероятностей в цепях Маркова Вестник Ленинградского Университета.
18. *В. И. Романовский.* Дискретные цепи Маркова, Гостехиздат, 1949.

Notă. In bibliografie, în afară de lucrarea lui A. Rényi nu am citat decît cărți. Citarea memoriilor ar fi constituit o listă cu mult prea incompletă și inegală.

Scanned with OKEN Scanner

ERATĂ

Pag.:	Rîndul:	In loc de:	Se va citi:	Din vina:
	7 sus	$= \sum_{k=1}^{m} p_{\cdot k} b_k$	$= \sum_{k=1}^{n} p_{\cdot k} b_k$	Editurii
216	3 sus	$\ldots q^{\beta}(1-p)^{n-\beta}$	$\ldots {}_1^{\beta}(1-q)^{n-\beta}$	Editurii
435	7 sus	$M_R\left(\frac{1}{\eta^4}\right) \geqq L^2 \cdot$	$M_{\bar{R}}\left(\frac{1}{\eta^4}\right) \leqq L^2 \cdot$	Editurii
685	4 jos	$\ldots e^{-\frac{(2k+1)^2\pi^2}{8}\frac{1}{\alpha a^2}}$	$\ldots e^{-\frac{(2k+1)^2\pi^2}{8}\frac{1-\alpha}{\alpha a^2}}$	Tipografie

Scanned with OKEN Scanner

Control științific : Traian Lișcu.

Responsabil de carte : Pop Avram
Tehnoredactor : Marcus Fany
Corector responsabil : Băluș Luiza

Dat la cules 20.10.1957. Bun de tipar 14.01.1958. Hîrtie cărți școlare, de 65 g/m², 61×86/16. Coli editoriale 39,50. Coli de tipar 45,75. Comanda P. R. 3417 A. 896.
Indicele de clasificare pentru bibliotecile mari 519.2.
Indicele de clasificare pentru bibliotecile mici 519.

Tiparul executat sub comanda nr. 6540 la Întreprinderea Poligrafică Timișoara, str. Popa Șapcă nr. 8. — R.P.R.

E. 14 662

Scanned with OKEN Scanner